UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC) - BARCELONATECH

FACULTAD INFORMÁTICA DE BARCELONA (FIB)

MÁSTER EN INGENIERIA INFORMÁTICA

Trabajo de fin de máster

Implementación del análisis de trayectorias en datos de procesos discretos en KLASS

Autor:
Ricardo Gregorio Calderón Valdiviezo

Directora:
Karina Gibert Oliveras

Departamento:
Estadística i Investigació Operativa

Fecha de presentación:
Octubre de 2019

Agradecimientos

Agradezco a mis padres y hermanos por el apoyo constante en las decisiones que he tomado durante lo largo de este Máster, a mi directora de Tesis Dra. Karina Gibert, por compartir sus conocimientos y por la paciencia que ha tenido durante todo el desarrollo de esta Tesis. A mis amigos que me han apoyado durante el máster con los que se han compartido horas de estudios

Resumen

La presente tesis de máster tiene como objetivo principal añadir una nueva funcionalidad en KLASS, el análisis de trayectorias en datos de procesos discretos. Se toma como punto de partida la clasificación basada en reglas por Estados (ClBRxE), cuyo marco teórico fue objeto de investigación y aplicación posterior y se publicó en [Gibert, García-Rudolph, Curcoll, Soler, Pla, Tormos 2009] y en [Gibert, Rodríguez-Silva, Rodríguez-Roda, 2010]. La tesis desarrolla un análisis de trayectorias siguiendo los pasos que se indican en [Gibert, Rodríguez-Silva, Rodríguez-Roda, 2010], para facilitar la interpretación de los resultados del clustering en procesos dinámicos

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción y objetivos

1.1 Introducción

Con el avance de la tecnología y de las técnicas de adquisición de conocimiento, se ha dado lugar al manejo intensivo de datos dentro de todos los ámbitos. Data Science [Gibert et al 2018] es una ciencia emergente que estudia cómo extraer el valor estratégico de los datos, desde la información proporcionada por varios medios como móviles, sensores, etc., y con la ayuda de otras ciencias como la matemática, la estadística, la inteligencia artificial, la informática… Puesto que se manejan dominios muy complejos y el análisis manual de los datos ya no es viable, se hace necesario desarrollar técnicas o herramientas que ayuden al proceso e interpretación de los datos de una forma sencilla.

Los métodos de minería de datos juegan un papel clave en el proceso de data science, así como la aportación de los expertos a lo largo de todo el proceso es fundamental también.

Sin embargo, para que realmente se pueda extraer valor de los datos, no basta con el análisis y los modelos que se generan a partir de la minería de datos, sino que es muy importante el postprocesado de los resultados del data mining.

Encontrar herramientas potentes e intuitivas de comunicación de resultados al experto es crítico, las decisiones que un experto pueda tomar van a depender de la interpretación que se logre hacer de los resultados del análisis. Así, si los resultados no son expresados de una forma clara y comprensible para un experto no necesariamente familiarizado con la tecnología, es posible que del análisis no se derive ningún impacto real en el dominio de aplicación. Herramientas como el diagrama de Barras, diagrama de BoxPlot se encuentran entre las visualizaciones más elementales de datos crudos que ya [Tukey 1977] recomendó como mecanismos de comprensión de datos. Más recientemente, herramientas más sofisticadas como el Class panel graph [Gibert, et al 2008a], el TLP [Gibert, Conti 2015] o la conceptualización CCEC [Gibert, 2014] ayudarán en la interpretación de los resultados

procedentes de un análisis de clasificación automática, sea cual sea el método de clasificación que la ha generado.

Cuando se analizan procesos dinámicos se ha de representar una realidad compleja, debido a que además de las habituales interacciones entre variables que se caracterizan en un clustering estático, existe una componente temporal o espacial (o ambas), que interviene en la evolución de los comportamientos y la síntesis simbólica de las herramientas de comunicación de los patrones descubiertos es más compleja

Klass es un software de soporte a ciencia de datos, el cual contiene una serie de componentes que permiten representar de forma visual el comportamiento del fenómeno de un estudio, a través de técnicas como: Clasificación, Interpretaciones basadas en boxplots o Caracterización Conceptual por Condicionamientos Sucesivos, Semáforos, Paneles de clases (Class panel graph) entre otras. El desarrollo de este software ha sido fruto de muchos años de investigación en la Universidad Politécnica de Cataluña bajo la dirección de la Dra. Karina Gibert directora de esta tesis.

Esta tesis de máster, está enfocada a ampliar la funcionalidad de KLASS con una nueva herramienta de visualización de patrones dinámicos, el diagrama de trayectorias, así como a integrar varias funcionalidades implementadas en Klass para tener como resultado final una representación visual de los comportamientos típicos de un proceso (trayectorias) que sigue un objeto, en base a las clases generadas por un método cualquiera de clasificación, si bien en esta tesis se utiliza la clasificación basada en reglas por estados (CIBRxE), que es una generalización de la clasificación basada en reglas [Gibert & Cortés 1998]

1.2 Objetivos

El objetivo de esta tesis aumentar la funcionalidad el sistema KLASS, con la Implementación del análisis de trayectorias en datos de procesos discretos en KLASS

Los objetivos específicos:

Incorporar a KLASS la herramienta de post-proceso asociada al método de clasificación basada en reglas por estados para completar la interpretación de resultados de este método [Gibert, Rodríguez-Silva, Rodríguez-Roda, 2010], hasta ahora realizada manualmente de forma externa a KLASS. Esto requerirá:

Generar trayectorias a partir de la tabla de frecuencias de los estados que participan.

Implementar estructuras que permitan almacenar las trayectorias generadas

Generar el diagrama de trayectorias como soporte gráfico para la interpretación de los resultados, siguiendo la filosofía de KLASS que genera todos los resultados como código fuente de LaTeX, que después se compila

Generar el reporte completo asociado al diagrama de trayectorias

Crear funciones para importar y exportar diagramas de trayectorias que permitan dar persistencia a los cálculos realizados para construirlos

Modificar el diseño de KLASS para que incorpore les clases y métodos necesarios para cubrir estas funcionalidades adicionales

Verificar el correcto funcionamiento de este nuevo módulo con juegos de datos disponibles, de distintas estructuras

Revisar la adaptación necesaria del módulo de Clasificación Basada en Reglas por Estados si fuere necesario

Verificar el correcto funcionamiento del nuevo módulo con varios casos reales

1.3 Estructura de la memoria

El presente documento este compuesto por la siguiente estructura

Capítulo 1: introducción y objetivos, contiene: contexto, objetivos, motivación y estructura del documento

Capítulo 2: metodología y antecedentes, contiene: proceso de clasificación por estados, introducción a java KLASS, desarrollo del proyecto.

Capítulo 3: caso de estudio de ejemplo aplicando el análisis de trayectorias en datos de procesos discretos

Capítulo 4: conclusiones

Glosario: definición de terminología utilizada en el desarrollo de este documento.

Bibliografía: incluye las referencias a toda la documentación utilizada en el desarrollo de este proyecto

Capítulo 2

Metodología y Antecedentes

La clasificación basada en reglas por estados (CIBRxE) fue desarrollada en la versión 8 de KLASS en el año 2009 como parte de proyecto de Tesis [García Rudolph, 2009]. El marco teórico de la ClBRxE fue asimismo objeto de investigación y aplicación posterior y se publicó en [Gibert, Rodríguez-Silva, Rodríguez-Roda, 2010]. Este método descansa sobre la clasificación basada en reglas (ClBR) [Gibert & Cortés 1998] que es un método mixto de clasificación automática que introduce conocimiento a priori como sesgo semántico del proceso de clasificación y que constituye el corazón de la herramienta KLASS, desarrollada en el departamento de Estadística e Investigación Operativa bajo la dirección de la Dra. Karina Gibert, directora de esta tesis de máster.

La ClBRxE es una generalización de la ClBR a la identificación de patrones de procesos discretos, representables como una cierta secuencia de estados y sobre los cuales hay información en la base de datos. A grandes trazos, la ClBRxE aplica la ClBR de manera separada a cada estado que tiene asignado un Proceso para después estudiar cómo se combinan las clases resultantes de cada estado. Dicho análisis se basa en una herramienta de post-proceso presentada en [García Rudolph, 2009] denominada diagramas de trayectorias, las cuales sintetizan las transiciones de los individuos entre los perfiles de cada estado del proceso. El diagrama de trayectorias está en la base para identificar los patrones de evolución del proceso dominantes y hasta el momento se realizaba manualmente en base a los resultados de la ClBRxE que se extraían de KLASS

En este proyecto de Tesis se diseñó e implementó el análisis de trayectorias en datos de proceso discretos, completando la integración de la metodología de ClBRxE en KLASS.

Adicionalmente, y debido a que las recientes ampliaciones de KLASS han introducido elementos que entraban en conflicto con el diseño original del módulo de ClBRxE, ha sido necesaria una revisión de dicho módulo y su adaptación a la actual estructura del sistema.

2.1 Proceso de clasificación basada en reglas por estados

2.1.1 Clasificación basada en reglas

La clasificación basada en reglas es un método de clasificación automática que incorpora elementos de gestión del conocimiento a priori como sesgo semántico a la construcción de las clases. Fue introducido por Karina Gibert [Gibert & Cortés 1998] a finales de los 90 y desde entonces se ha venido utilizando en la identificación de patrones en dominios reales de altos niveles de complejidad. El procedimiento es el siguiente

Dada una matriz de datos con $\mathrm{I} = \{\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_\mathrm{n}\}$ elementos en las filas, descritos por K variables $V_1, V_2, \dots, V_K$,

- Adquisición de conocimiento a priori: construcción de la base de conocimiento BC que cumpla con los objetivos del experto.

- Calcular la Partición inducida por las reglas P_R sobre I, evaluando R sobre $V_1, V_2, \dots, V_K$, incluyendo una clase residual C_0 que contendrá todos los objetos de I que no cumplen ninguna regla de BC

- Clasificar por separado cada elemento de P_R l excepto C0, es decir, para cada una de las clases inducidas por las reglas R de BC construir una clasificación local a dicha clase. Se propone aquí utiliza la clasificación ascendente jerárquica con el criterio de Ward y en caso de que haya variables numéricas y cualitativas simultáneamente utilizar la distancia mixta de Gibert [Gibert Cortes 1997]

- Integrar la clase residual a los prototipos de las clases inducidas por R. y clasificarlos de manera conjunta, recuperando un nuevo conjunto de datos con todo el dato de C0 y los prototipos de las clases inducidas por las reglas, obteniendo un nuevo dendrograma

- Recuperar la estructura jerárquica asociada a los prototipos e integrarla en el dendrograma obtenido en el paso anterior generando una clasificación global del conjunto $\mathrm{I} = (\mathrm{i}_1, \mathrm{i}_2, \dots, \mathrm{i}_\mathrm{n})$ con una jerarquía global (dendrograma) que integra todos los elementos de I en sus hojas

- Una vez realizada la clasificación y generado el dendrograma resultante, con ayuda del experto se procede a realizar la visualización y el corte para maximizar la ratio

de Calinski-Harabasz [Benzecri 1973]. El resultado es la partición P que asocia una clase a cada individuo de I

- Si se genera inconsistencias en la interpretación, se deberá de corregir de acuerdo con el conocimiento a priori que se posee, y volver a realizar la clasificación.

2.1.2 Clasificación basada en reglas por estados

La clasificación basada en reglas por estados (CIBRxE), es una generalización de la clasificación basada en reglas [Gibert & Cortés 1998] La metodología que se aplica para ClBRxE principalmente es el uso reiterado de CIBR separando las variables por estados. Ello requiere previamente de la introducción en el sistema de la información que permita identificar cuáles de las variables disponibles en la base de datos aportan información a cada uno de los estados del proceso y se sitúa dentro del ámbito de la gestión de la metainformación sobre el problema, aspecto poco abordable en los softwares de análisis de datos actuales.

Definiendo el Proceso como un conjunto de Estados o Subprocesos $S = (e_1, e_2, \dots, e_E)$, donde cada estado está conformado por $I = (i_1, i_2, \dots, i_n)$ observaciones descritas por $V_1, V_2, \dots, V_K$ variables y definiendo una base de conocimiento, en forma de reglas Lógicas R , se propone una síntesis de los pasos sugeridos para proceder :

- Calcular la Partición inducida P_R sobre I, evaluando R sobre $V_1, V_2, \dots, V_K$, incluyendo una clase residual C_0.

- Dividir el conjunto de variables según el estado al que pertenezca, donde $V_1^e, \dots, V_{Ke}^e$ es el subconjunto de variables del Estado e.

- Para cada Estado $e \in S$, $S = (e_1, e_2, \dots, e_E)$ se procede iterativamente:

    - Seleccionar de la matriz de datos las variables correspondientes al estado e, formando la submatriz de datos Xe ($V_1^e, \dots, V_{Ke}^e$)

    - Se obtiene $ID_e = \{i \in I: X_{i1}^e = X_{i2}^e = \cdots = X_{ike}^e = *\}$ el conjunto de objetos de I que en el estado e no tienen información para ninguna variable

    - Efectuar una clasificación basadas en reglas (CIBR) [Gibert & Cortés 1998] sobre I / ID_e con la matriz Xe utilizando P_R como partición inducida por R a cada estado (seguir los pasos del apartado anterior)

- Una vez realizada la clasificación y generado el dendrograma resultante, con ayuda del experto se procede a realizar la visualización y el corte para maximizar el ratio de Calinski-Harabasz [Benzecri 1973]. Sea P_e^* la partición de I obtenida en el estado e

- Valoración y etiquetado por el experto de las clases resultantes $C \in P_e^*$ haciendo uso de las herramientas de soporte como el Class Panel Graph [Gibert et 2005] y del cuadro de semáforos.

- Construir $P = (P_1, P_2, \dots, P_E)$ que constituye la clasificación de I según la información de cada uno de los estados del proceso.

- Realizar el análisis de las trayectorias más probable a partir de P. Para esto se deberán seguir los pasos del apartado 2.1.4

- Las trayectorias seleccionadas constituyen patrones dinámicos que describen la evolución del proceso en términos discretos y cualitativos

2.1.3 Definición de Trayectorias

Se define como trayectoria una τ en forma de secuencia de clases que describen el paso de cierto objeto por las diferentes etapas de un proceso. $\tau = (\ C_\tau^1, C_\tau^2, \dots, C_\tau^E\ )$, entendiendo que cada etapa del proceso se puede encontrar en un número finito de estados C_τ^{e}

Definición de Diagramas de trayectorias

Después de generar una clasificación por estados KLASS dispone de una lista de dendrogramas, uno para cada estado del proceso, que el usuario analizará para decidir el nivel del corte y obtener u conjunto de clases para cada estado y creará las correspondientes listas de variables cualitativas en la matriz de datos, indicando para cada individuo en que clase se sitúa para cada uno de los estados del proceso.

Dado un conjunto de trayectorias: $\tau = (\ C_\tau^1, C_\tau^2, \dots, C_\tau^E\ )$

Un diagrama de Trayectorias está definido como un grafo direccional compuesto por nodos y arcos, los nodos son representaciones gráficas de los estados de las distintas etapas del proceso, y los arcos de las trayectorias τ que se dan en el proceso.

Los Nodos están dispuestos en columnas y cada columna representa una etapa del proceso (que formalmente se representa como una variable cualitativa), y cada elemento representa los diferentes estados que puede presentar dicha etapa, que formalmente corresponderán a las distintas modalidades que puede tener la variable cualitativa que representa el estado en cuestión. Estas modalidades pueden presentarse en el diagrama de trayectorias ordenadas en función de un criterio especifico o seguir el orden de original de representación en la matriz de datos.

Los arcos del diagrama representan las transiciones entre estados de las distintas etapas de cada trayectoria y por tanto aportan una descripción cualitativa de lo que ocurre en procesos discretos.

Para construir un diagrama de trayectorias a partir de $\mathrm{P} = (\mathrm{P}_1, \mathrm{P}_2, \dots, \mathrm{P}_\mathrm{E})$ se han de seguir los siguientes pasos [Gibert, Rodríguez-Silva, Rodríguez-Roda, 2010]:

- Definir un espacio de estados formado por todas las clases de todas las particiones de P, dispuestos en columnas de acuerdo con dichos estados.

- Identificar las trayectorias que ocurren en P (el conjunto de trayectorias posibles de P es el producto cartesiano de sus componentes $P_1 \mathrm{x} P_2 \mathrm{x} \dots \mathrm{x} P_E$ pero en un caso real no se observan todas las combinaciones posibles, sino solamente un pequeño subconjunto de las mismas)

- Calcular la frecuencia observada con que aparecen cada una de las trayectorias identificadas en el apartado anterior

- Definir las trayectorias que se desean visualizar (con referencia a algún criterio…. Las más frecuentes, las más infrecuentes…. Relacionado con los objetivos del análisis)

- Construir diagrama de trayectorias definitivo a visualizar formado por el espacio de estados detallado en el primer punto y arcos que representen las transiciones entre clases de estados sucesivos de las trayectorias seleccionadas

En realidad, el diagrama de trayectorias se concibe originalmente para visualizar los patrones correspondientes a procesos dinámicos de cierta complejidad. Sin embargo, la estructura del elemento de partida sobre el que se construyen $\mathrm{P} = (\mathrm{P}_1, \mathrm{P}_2, \dots, \mathrm{P}_\mathrm{E})$ corresponde a la de un vector de variables cualitativas y por ello, el diagrama de trayectorias tiene una acepción más general que consiste en visualizar las interacciones entre varias variables cualitativas. Es por ello que se ha optado por implementar esta herramienta para un contexto general basado en cualquier subgrupo de variables

cualitativas y cuando se utilice para visualizar las trayectorias formadas por las particiones de los estados de un proceso, aportará una interpretación adicional en términos de dinámica de procesos.

2.2 Introducción a Java-Klass

2.2.1 Introducción a Klass

KLASS es un sistema interactivo con 30 años de desarrollo controlado, liderado por Karina Gibert, como parte de su tesis de licenciatura en su primera versión [Gibert, 1991] y posteriormente como parte de su tesis doctoral [Gibert,1995], fue propuesto a la Facultad Informática de Barcelona (FIB) donde fue diseñado y Desarrollado. Este está compuesto de un conjunto de funcionalidades que sirven de ayuda para los especialistas en el área de minería de Datos, facilitando la interpretación de los datos.

Fue desarrollado inicialmente en LISP sobre UNIX, siendo migrado a java posteriormente para aprovechar aprovechando entre otras cosas la portabilidad que brinda este lenguaje sobre el anterior, ya que puede ser ejecutado en diferentes sistemas operativos, además del costo cero de las licencias en comparación con LISP, entre otras ventajas.

Con el paso del tiempo Java-Klass ha ido añadiendo y mejorando funcionalidades a su versión Original, como parte de tesis de Grado, de Master o Doctorales, además como parte de proyectos de varias materias dictadas en la Universitat Politécnica de Catalunya (UPC) o la Universitat Illes Balears (UIB). Además, es utilizado en el ámbito investigativo, como parte de proyectos y estudios en diferentes áreas de la investigación.

Arquitectura de KLASS

La arquitectura de Klass está conformada por lo siguiente:

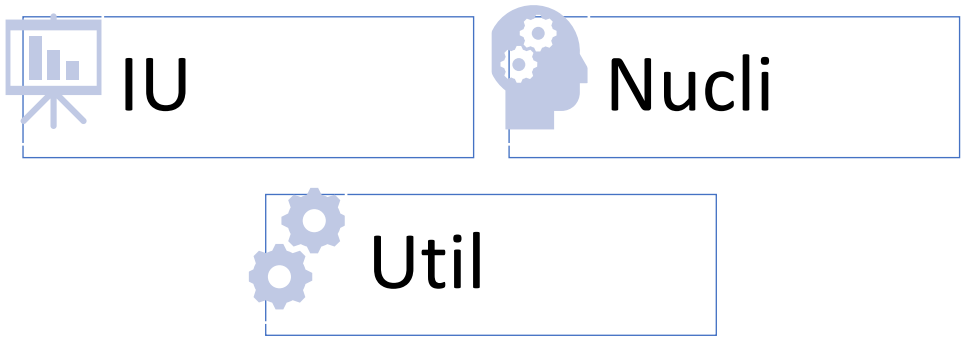


Figura 1 Arquitectura KLASS

Dónde UI está conformado por la interfaz de la plataforma, NUCLI es la parte fundamental del sistema, y por último UTIL que son librerías componentes desarrollados en KLASS para el perfecto funcionamiento de todo el entorno.

Para esta tesis de máster se trabajó en todo el componente, para cumplir con todos los requerimientos definidos en la planificación.

2.2.2 Cronología de Java-KLASS

A continuación, un breve resumen de cómo ha evolucionado Java-Klass hasta su última versión:

Feb. 1991 KLASS v0. Tesina Karina Gibert. \KLASS. Estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades". Classifica matrius de dades heterogènies amb la distància mixta.

Nov. 1994 KLASS v1. Tesi Karina Gibert. “L'us de la informació simbòlica en l'automatització del tractament estadístic de dominis poc estructurats". És una ampliació de KLASS v0. Incorpora la classificació basada en regles. [Gibert 94]

Jul. 1996 KLASS v1.1. PFC Xavier Castillejo. Incorpora a KLASS.v1 una interfície de finestres independent amb un sistema que facilita l'ús de KLASS des de SUN i des de PC a usuaris que desconeixen Lisp i UNIX. Anomenarem xcn.KLASS al nucli Lisp d'aquesta nova versió i xcn.i a la interfície C.

Oct. 1997 jj.KLASS. PFC Juan José Márquez i Juan Carlos Martín. Incorpora a la versió KLASS.v1 noves opcions per al tractament de dades mancants, la possibilitat de treballar amb objectes ponderats i implementa un test no paramètric de comparació de classificacions.

Set. 1999 KLASS v1.2. PFC Xavier Tubau (versió β). Incorpora a la versió xcn.KLASS el mòdul de comparació de classificacions de jj.KLASS, la mètrica mixta de Ralambondrainy i prepara la formulació de tres més per la seva posterior implementació. Anomenarem xt.KLASS al nucli Lisp d'aquesta nova versió i xt.i a la interfície C associada.

1999-2000 KLASS+ v1. PFC Sílvia Bayona. Fusió definitiva de la versió xt.KLASS amb jj.KLASS. Incorpora a més un mòdul no d'anàlisi descriptiva de les dades, també de les classes resultants, reorientant KLASS cap a un propòsit més general i menys especialitzat. Anomenarem sbh.KLASS al nucli Lisp d'aquesta nova versió i sbh.i a la interfície C associada.

2000-2002 KLASS+ v2. PFC Josep Oliveras. Afegeix a sbh.KLASS les mètriques mixtes pendents (Gower, Gowda-Diday i Ichino-Yaguchi). Anomenarem joc.KLASS a aquesta nova versió.

2000-2003 jr.KLASS+. Tesi doctoral Jorge Rodas. Integra KLASS+ v.2 i Columbus, que s'introdueix més endavant.

2000-2003 Recerca Anna Salvador i Fernando Vázquez. Desenvolupament de CIADEC, que s'introdueix més endavant.

2002-2003 Java-KLASS v0. PFC Ma del Mar Colillas. Versió Java del mòdul d'anàlisi descriptiva i integració amb CIADEC i Columbus.

2003-2005 Java-KLASS v0.22. Col•laboració amb Mar Colillas. Ampliació de l'anàlisi descriptiva i introducció d'eines de gestió de dades (definició d'ordenacions en els informes, possibilitat de varies matrius d'objectes en el sistema simultàniament, canvi de matriu activa).

2005-2006 Java-KLASS v1.0. Col•laboració amb Mar Colillas. Inclou la lectura i visualització de dendogrames aïllats, així com la generació de particions a partir d'ells.

2006-2007 Java-KLASS v2.0. PFC José Ignacio Mateos. Ampliació de Java-KLASS amb un mòdul de càlcul de distàncies per diferents tipus de matrius de dades, incloent les que combinen informació qualitativa i quantitativa, tractament de missings i creació de submatrius.

2006-2007 Java-KLASS v3.0. PFC Roberto Tuda. Inclou un mòdul de classificació automàtica per mètodes jeràrquics, utilitzant totes les distàncies implementades a la v2.0 i una opció per a estudiar agregacions d'objectes pas a pas. Es crea la opció de poder seleccionar el directori de treball per defecte. Se li agrega la opció de afegir-hi i desar objectes amb pes.

2006-2007 Java-KLASS v4.0. PFC Laia Riera Guerra. Introducció, gestió i avaluació de Bases de Coneixement. Ampliació de Java-KLASS amb un mòdul de transformació de variables que permet discretitzacions, recodificacions i càlculs aritmètics amb variables numèriques. Per últim, aquesta versió inclou la definició de submatrius via filtres lògics sobre els objectes, l'edició de metainformació de les variables de la matriu, eliminació de variables i importació de fitxers en format .dat estàndard.

2007 Java-KLASS v5.0. PFC Andreu Raya. Inclou la classificació condicionada, la classificació basada en regles i funcionalitats de divisió de la base de Dades i de gestió d'arbres de classificació (o dendogrames) associats a les diferents matrius de dades.

2007 Java-KLASS v6.0 Treball d'investigació Tutelada Alejandro García. Classificació basada en regles exògena. Intenacionalització i localització de a tres idiomes (Català, Anglès i Castellà). Fusió de matrius.

2008 Java-KLASS v6.4. Treball de Màster Alfons Bosch Sansa, Patricia García Giménez, Ismael Sayyad Hernando. Boxplot-based discretization, Boxplot-based Induction rules.

2008 Tesi doctoral Alejandra Perez. Caracterització per condicionaments successius, metodologia que indueix automàticamwnr a conceptes associats a les classes descobertes.

2008 Tesi doctoral Gustavo Rodríguez. Classificació basada en regles per estats que permet anàlisi de sistemes dinàmics.

2008: Java-KLASS v7.0: TrT Alejandro García Rudolph. Fusió de matrius i gestió de variables actives.

2009: Java-KLASS v8.: Tesi de màster d'Ester Lozano. Criteris Best Local Concept and no close world assumption del CCEC. PT Alejandro García Rudolph. Classificació basada en regles per estats.

2010: Java-KLASS v8.1: Practica SISPD. Narcis Maragall. Boxplot Based Induction Rules.

2012: Java-KLASS v8.6: Practica SISPD. Pau. metodología CCEC.

2012: Java-KLASS v9: Practica SISPD. Marco Villegas. Criteris CCEC.

2013 Java-KLASS v10: Practica SISPD. Emili Boronat. Traffic Light Panel.

2014: Java-KLASS v11: Projecte final de Carrera Enginyeria Informàtica FIB. Sheila Mollà. DBSCAN, OPTICS, 3D Visualization.

2014: Java-KLASS v12: Practica SISPD. Jonathan Moreno. Optimització d'expressions lógiques.

2015 Java-KLASSv15: Practiques IKPDI+SISPD Sergio Santamaria i Daniel Gibert et alt pràctiques Gestió d'ONTOLOGIES, distàncies semàntiques. Classificació basada en ontologies.

2016 Java-KLASSv16: TFG Valerio Di Matteo (U. La Sapienza, Roma, Italy). Mostreig i Escalabilitat: Generació de variables aleatòries, extracció de mostres aleatòries sobre la matriu de dades, k-Nearest Neighbour, CURE.

2016 Juny Java-KLASSv17: TM David Canudes + practiques IKPD des2015: Gestió termòmetres+automatització de TLPs.

2016 nov Java-KLASSv18 pràctiques IKPD: Implementació de TLPs anotats. Primeres infrastructures per gestionar variables multivaluades (desplegament i concatenacions).

2018 mar Java-KLASSv18.2: TM Luis Daniel Pérez Tamayo: Gestió de variables multivaluades i consolidació treball anterior

2018 April Java-KLASSv18.3: TM Johnny Avila: Termòmetres qualitatius, connexió amb semàfors

2018 Maig Java-Klass v18.4: TM Carlos Luis Jordán: reorganització de tots els mètodes d'inducció de conceptes i metodologia de resolución de conflictes en bases de coneixement

2018 Juny Java-Klass v19: TFG Lavanya Mandadapu: models predictius, gestió de dummies, aggregació de variables a partir de factors.

2019 January Java-Klass v20: TFM Juan de los Reyes Piedra: MIMMI method.

2019 June Java-KLASS 21.0: TFM Andres Bermudez: Clasification based on rules with ontologies

2019 June Java-KLASS 21: TFM Javi Vásquez: Date variables management

2019 October Java-KLASS 22: TFM Ricardo Calderón: Trajectories Diagram creation and management

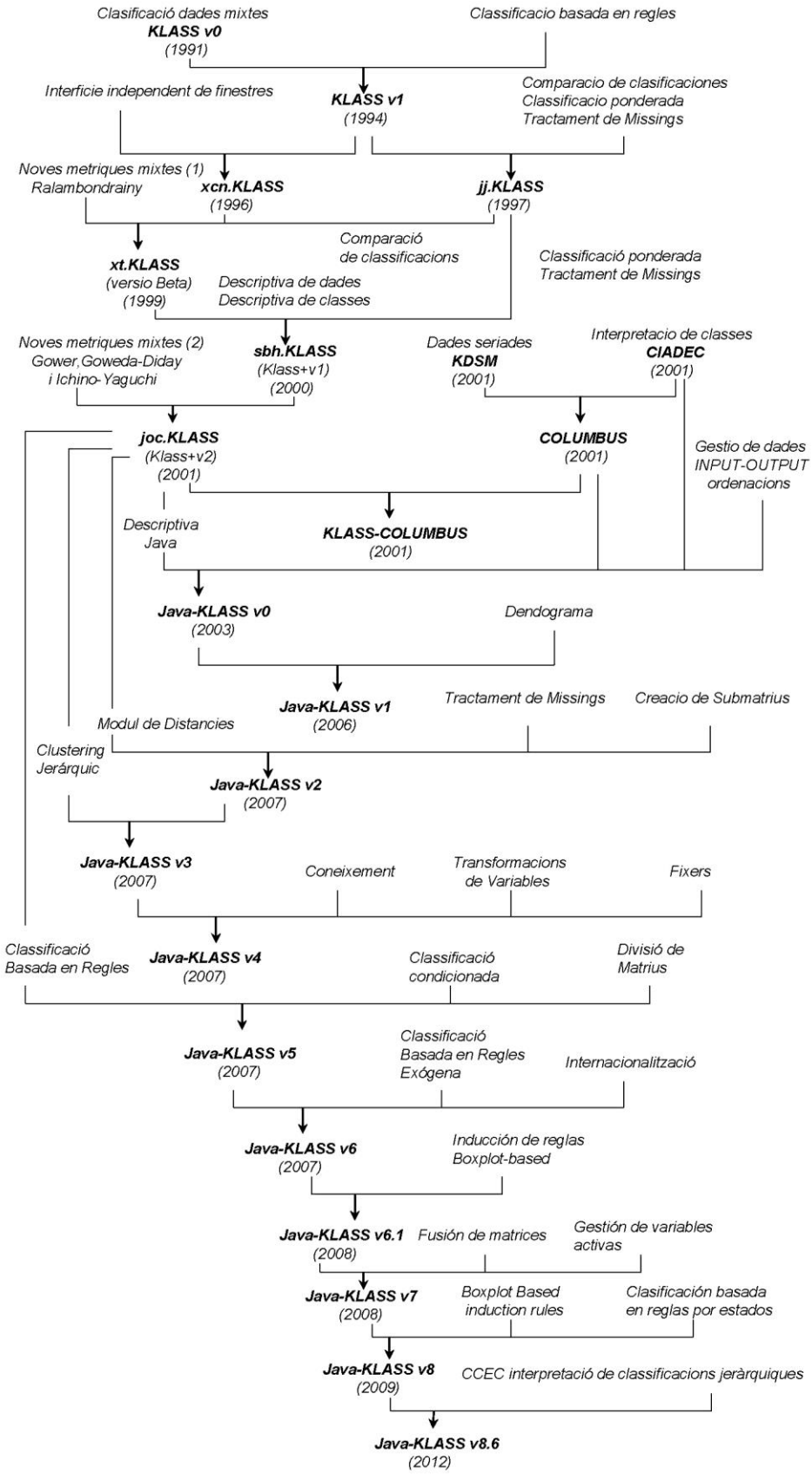


Figura 2 Cronología Java-KLASS (parte 1)

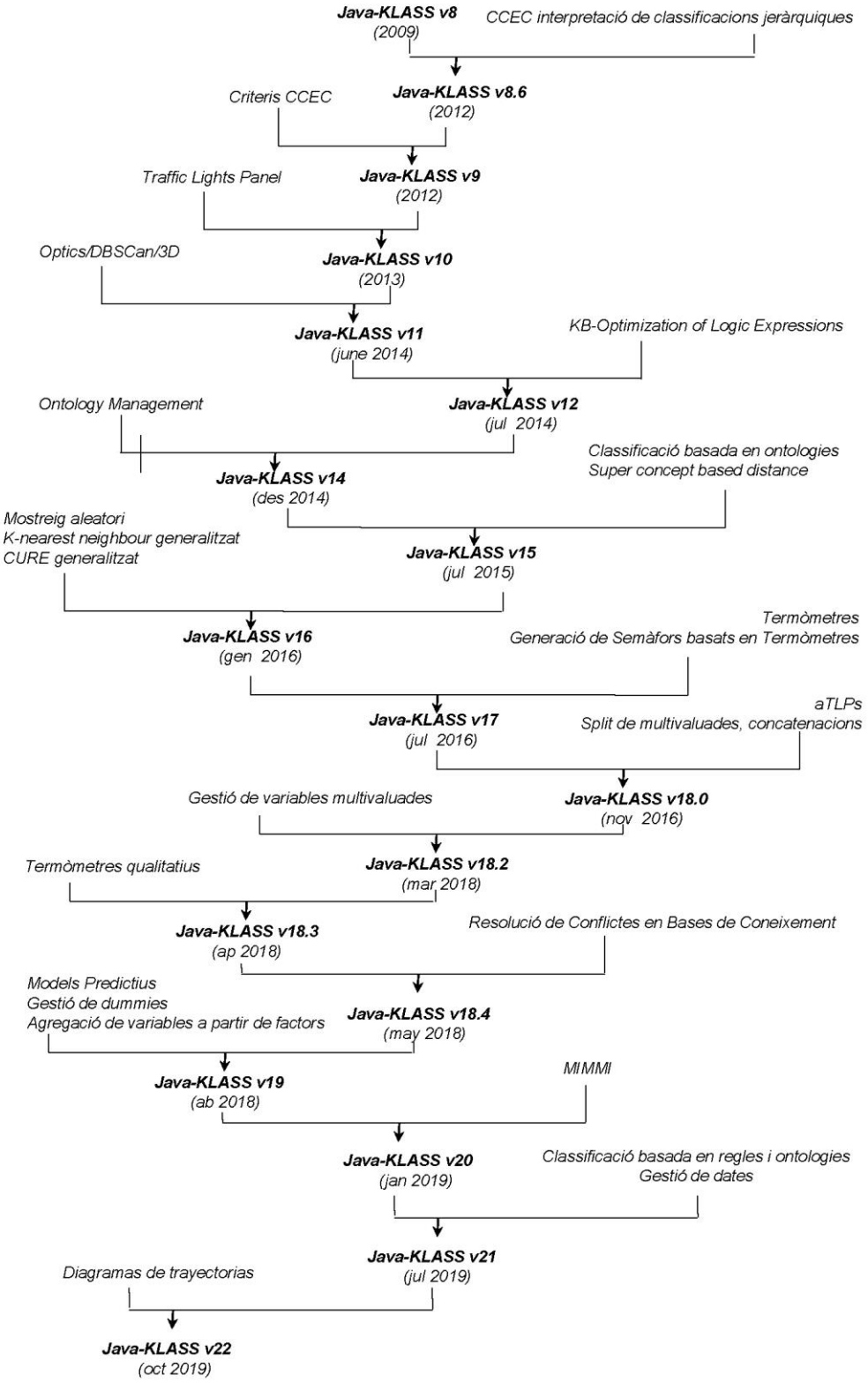


Figura 3 Cronología Java-KLASS (parte 2)

2.2.3 Funcionalidades de Java-KLASS

En el presente apartado se presentará un listado con las funcionalidades del sistema Java-KLASS en versión actualizada (v22):

- Representación de matrices de datos, variables cualitativas, cuantitativas, semánticas y manejo de metadatos
- Selección de variables e individuos basados en criterios definidos por el usuario para generar submatrices. Generación de submatrices por muestreo aleatorio bajo diferentes esquemas
- Recodificación o discretización de variables y generación de variables nuevas.
- Gestión de bases de conocimiento.
- Gestión y visualización de ontologías
- Gestión y visualización de termómetros
- Estadística Descriptiva univariante, bivariante y trivariante de los datos y de distribuciones condicionadas (CPGs).
- Visualización 3D.
- Análisis dinámico [Gibert et al., 2010a].
- Cálculo de distancias con métricas de distintas familias como: Euclídea, Métrica del valor absoluto, Minkovski, mixta de Gibert [Gibert et al. 2005], Ralambondrainy, Gower, Gowda-Diday, Ichino-Yaguchi, mixta de Gibert generalizada, Chi-cuadrado, Hamming generalizado, s seda-based distance.
- Custering automático con métodos jerárquicos clásicos, basados en reglas, en ontologías, con métodos basados en densidades como BDSCAN o OPTICS [Molla Santiago, 2014].
- Interpretación de clases vía TLP, IRBP [Gibert et al., 2012b, Gibert et al., 2013] [Gibert and Conti, 2016, Gibert et al., 2008a] y conceptualmente CCEC [Gibert 2014]

- Métodos de interoperabilidad.

- Gestión de sistemas heterogéneos que incluye información numérica, cualitativa, semántica, BC, ontologías, termómetros

2.3 Desarrollo del proyecto

2.3.1 Especificaciones

Este proyecto de Tesis se centra en el diseño e implementación del análisis de trayectorias en datos de proceso discretos, completando la integración de la metodología de ClBRxE en KLASS.

Adicionalmente, y debido a que las recientes ampliaciones de KLASS han introducido elementos que entraban en conflicto con el diseño original del módulo de ClBRxE, ha sido necesaria una revisión de dicho módulo y su adaptación a la actual estructura del sistema.

Es necesario tener en consideración los siguientes requerimientos:

1. Se debe de Garantizar el correcto funcionamiento de las demás funcionalidades de KLASS.

2. Revisión del módulo de definición de proceso, realizando una correcta asignación de estados al proceso creado, por medio de estructuras de almacenamiento.

   Se creará proceso y cada proceso debe de contener una lista de estados asociada, y cada estado deberá de contener una lista de variables.

3. Adaptación del módulo de clasificación por estados a la nueva estructura diseñada para el manejo de procesos y estados. Para el proceso seleccionado se debe de ejecutar la CIBRxE para cada estado que tenga el proceso, creando un dendrograma por clasificación de estado.

4. Realizar el cálculo de trayectorias de una lista de variables cualitativas para esto se usará el algoritmo MPT (mencionado en el capítulo 2.1.1), para esto se deberá de realizar los siguientes pasos:

- Concatenar las variables cualitativas seleccionadas, creando una nueva columna en la matriz de datos con la que se está trabajando.

- Realizar un análisis descriptivo de la nueva columna creada para obtener la tabla de frecuencias, con las frecuencias relativas y absolutas.

- Crear una estructura de almacenamiento (lista), que contendrá la información de cada nodo.

5. Crear una nueva opción dentro de Klass Panel denominado Diagrama de trayectorias, que permita la creación del diagrama de trayectoria calculado, y deberá de contener:

    - Se debe de permitir seleccionar las variables que se desea graficar a partir de la lista de variables cualitativas que se encuentran en la matriz de datos.

    - Se debe de permitir visualizar los valores asignados a cada variable mediante un botón.

    - Permitir cambiar el nombre de la nueva variable, garantizando que dicho nombre no generar conflicto con variables utilizadas internamente por KLASS

    - Permitir al usuario decidir si conserva o no la variable creada, cada vez que ejecuta el proceso.

    - Permitir al usuario decidir cuantas trayectorias desea visualizar, mediante una opción además se debe de validar que el número de trayectorias por representar no debe de ser mayor que el número total de trayectorias generadas

    - Permitir al usuario decidir la forma de pintar los arcos del diagrama de trayectoria, en la que pueda escoger entre la forma estándar de pintar los arcos o la forma normalizada, donde se normaliza la frecuencia de cada arco con respecto a la frecuencia mayor.

- Generación de Reporte en codificación LaTeX, mostrando el diagrama de trayectoria generado, la tabla de frecuencia y diagrama de barras de la variable de trayectoria creada, y la tabla de frecuencia y diagrama de barra de cada variable que conforma la variable de trayectoria por separado

6. Reutilización métodos de las funcionalidades ya implementadas en KLASS en los casos que se necesiten, realizando llamadas a los mismos con la parametrización necesaria para la implementación de la nueva funcionalidad

7. De requerir nuevos objetos, clases java, interfaces, se debe de conservar los estándares propios del sistema

8. Creación de manual de usuario para las nuevas funcionalidades del sistema.

2.3.2 Definición de procesos

En versiones anteriores de Java-Klass, se desarrolló la funcionalidad para definición de proceso, el cual consiste en diseñar una lista de procesos con sus respectivos estados, cada estado debe tener asignado una lista de variables. Sin embargo, al ejecutar esta funcionalidad en escenarios donde había agregar más de un proceso asociado a la misma matriz de datos, existían conflictos al vincular cada proceso con su cada uno de sus estados y sus variables asociadas

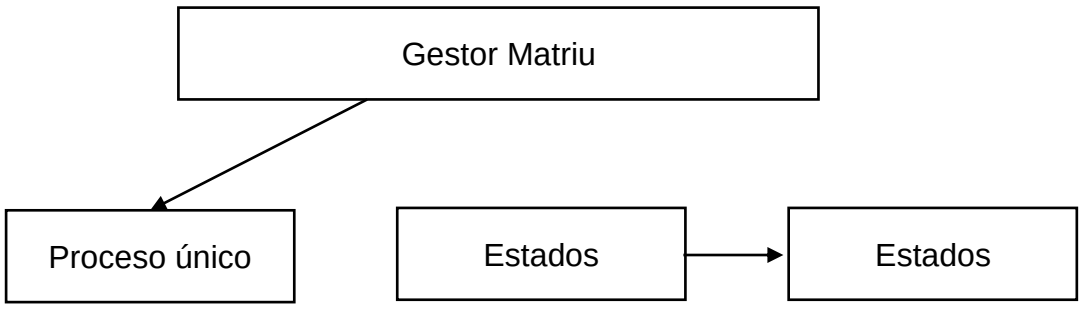


Figura 4 Definición de proceso

Este diseño presentaba la limitación de que se podía definir un único proceso asociado a una matriz de datos con una única lista de variables para cada estado del proceso. Si

el usuario pretendía asociar diferentes procesos a una misma base de datos se presentaba una disfunción por la que KLASS confundía todas las definiciones en un único proceso y generaba análisis incorrectos.

Sin embargo, se han presentado diversas aplicaciones reales en las que tiene pleno sentido definir varios procesos independientes que se puedan querer analizar sobre una misma base de datos y se ha planteado aquí una ampliación del diseño que pueda tratar con este escenario correctamente.

A cada base de datos se le añadirá una estructura de almacenamiento (Lista) de procesos y cada proceso a su vez tenía asignada una estructura de almacenamiento (lista) de Estados

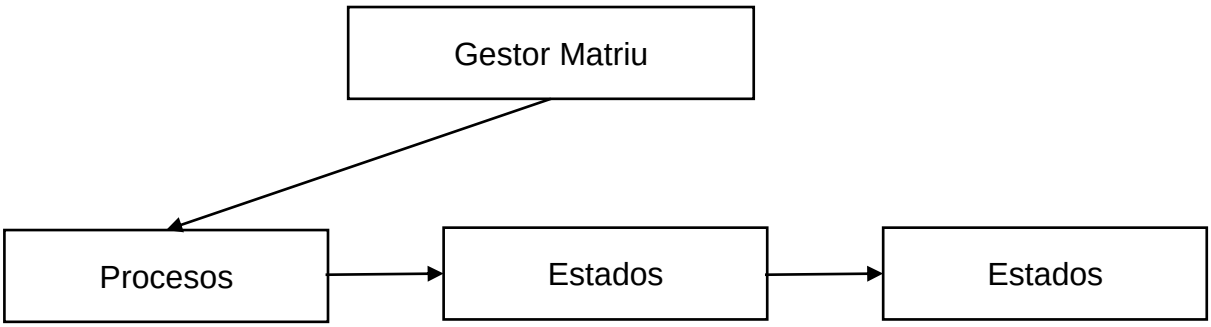


Figura 5 Definición de procesos con listas

De esta forma en el panel de clasificación por estados (que incluye a su vez una generalización de la ClBRxE permitiendo utilizar distintos métodos de clustering como base), se mostrará de manera correcta todos los procesos definidos sobre la matriz de análisis y la lista de estados que pertenecen al proceso seleccionado, procediendo a realizar la clasificación por estados de forma correcta.

En versiones anteriores, para poder realizar una clasificación por estados, Klass tomaba un proceso y consultaba cada uno de sus estados, las que contienen una lista de modalidades [García Rudolph, 2009]. Sin embargo, al no existir un vínculo entre los estados y los procesos producía un error al clasificar ya que un estado no se encontraba en ningún proceso específico, Implementando la nueva estructura de definición de procesos se resuelve esta disfunción y se realizar la clasificación por estados de forma correcta generando N dendrogramas donde N corresponde al número de estados del proceso seleccionado y se almacenan en la lista de dendrogramas asociadas a la matriz de datos.

2.3.3 Diagrama de trayectorias

Estas variables de estado nuevas constituyen el punto de entrada para realizar el diagrama de trayectorias como se lo describe en el apartado 2.1.3. Así, se ha dispuesto un panel que aporta una funcionalidad algo más general y permite construir diagramas de trayectorias sobre la base de conjuntos de variables cualitativas cualesquiera, disponibles en la Base de datos, lo cual, a la larga, además de construir diagramas de trayectorias sobre procesos, permitirá estudiar las distribuciones conjuntas de varias variables cualitativas en otros contextos.

El usuario pues seleccionará las variables de estado con las que representar el diagrama de trayectorias de entre todas las variables cualitativas disponibles en la matriz de datos.

KLASS aportará información sobre el número total de trayectorias distintas que ocurren en la base de datos y cuantificarás sus presencias. Ante la naturaleza combinatoria de la variable Trayectoria, el usuario podrá definir el número de trayectorias que desea representar en la figura

El proceso genérico es el siguiente:

1. Obtener lista de variables cualitativas
2. Obtener la lista ordenadas de modalidades de cada una de las variables.
3. Identificar las trayectorias que ocurren en los datos observados
4. Obtener la tabla de frecuencia de la variable de trayectorias
5. Ordenar por frecuencia para determinar los caminos con mayor frecuencia
6. Generar reporte LaTeX

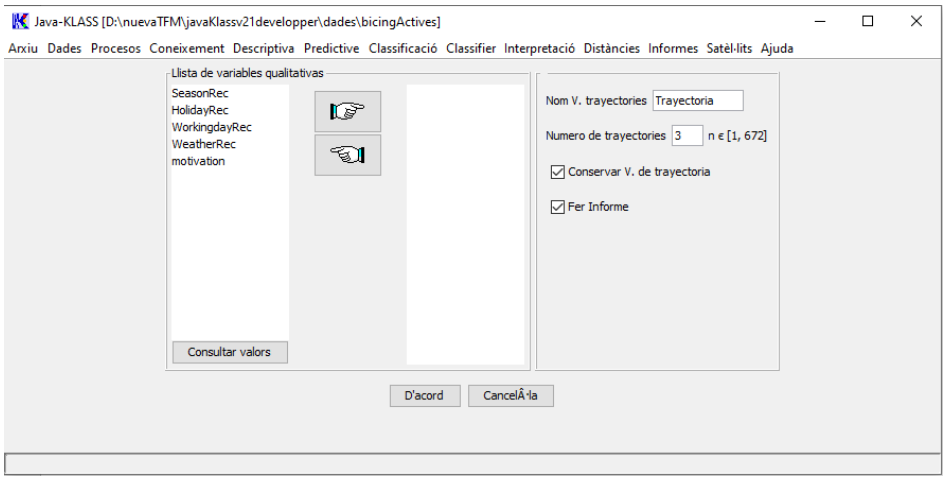


Figura 6 Interfaz Panel Diagrama trayectorias

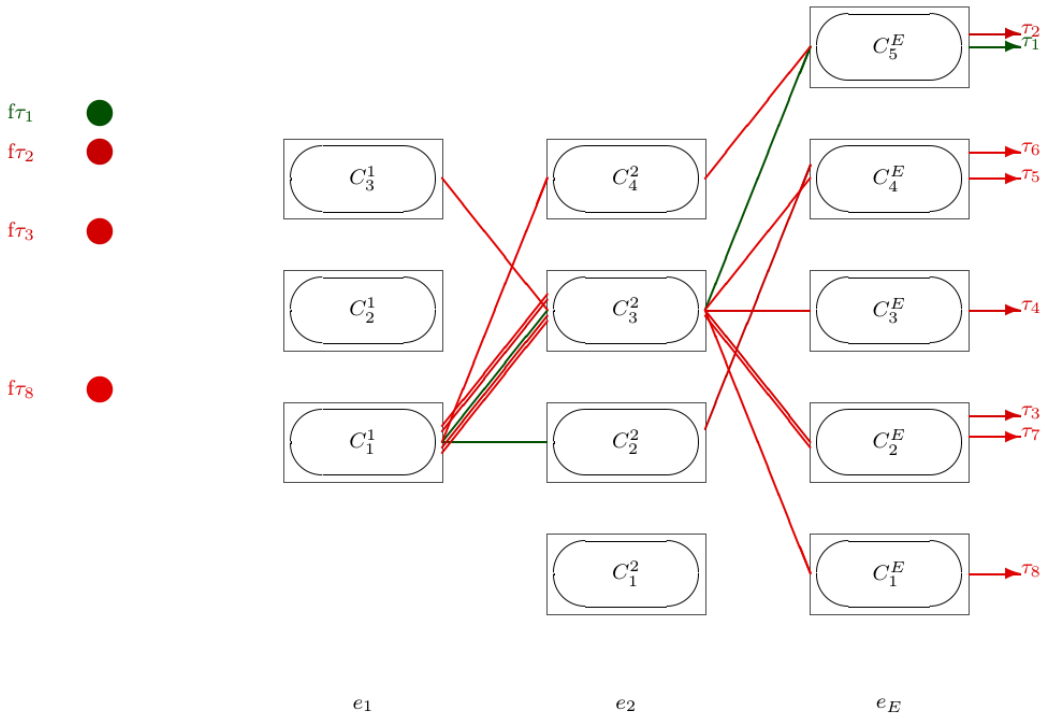


Figura 7 Ejemplo Diagrama de trayectorias

2.3.4 Diseño e implementación

En esta sección se detallará el trabajo realizado durante el desarrollo de esta tesis, en primer lugar, se tratará el tema de la definición de proceso, describiendo el funcionamiento previo a la implementación realizada, la mejora de la clasificación por estados.
Posteriormente detallaremos la implementación del análisis de trayectorias en datos de proceso discretos, los mecanismos usados, clases, estructuras creadas, y luego se detallará la interface creada y como debe de ser manipulada por el usuario para la generación del diagrama de trayectorias.

2.3.4.1 Definición de procesos

La definición de procesos es una funcionalidad de KLASS, que permite definir un proceso que está compuesto por un conjunto de estados y cada estado a su vez se describe a partir de varias variables. La definición de procesos sobre bases de datos en KLASS es una funcionalidad que se debe utilizar previamente a la realización de una clasificación basada en reglas por estados, donde es necesario seleccionar un proceso con sus respectivos estados. Sin embargo, en la versión 21 de KLASS, no hay una

asignación correcta los estados a un proceso, lo que provoca un conflicto al momento de realizar la clasificación, donde permitía crear varios procesos, pero solo el último proceso creado era el que se le asignaban los estados y procesos creados anteriormente.

Es por esto por lo que se diseñó e implemento una mejora, esta consistía en:

Creación de estructuras de almacenamiento (Listas) que van a contener los procesos que han sido definidos.

Creación de estructuras de almacenamiento (Listas) que contendrán los estados de cada proceso

Manteniendo estructuras de almacenamiento (Listas) de variables que eran asignadas a cada estado creado.

De esta forma cada proceso tendrá asignado una lista de estados y cada estado tendrá asignado una lista de variables, teniendo de esta forma una correcta relación proceso-estado-variable.

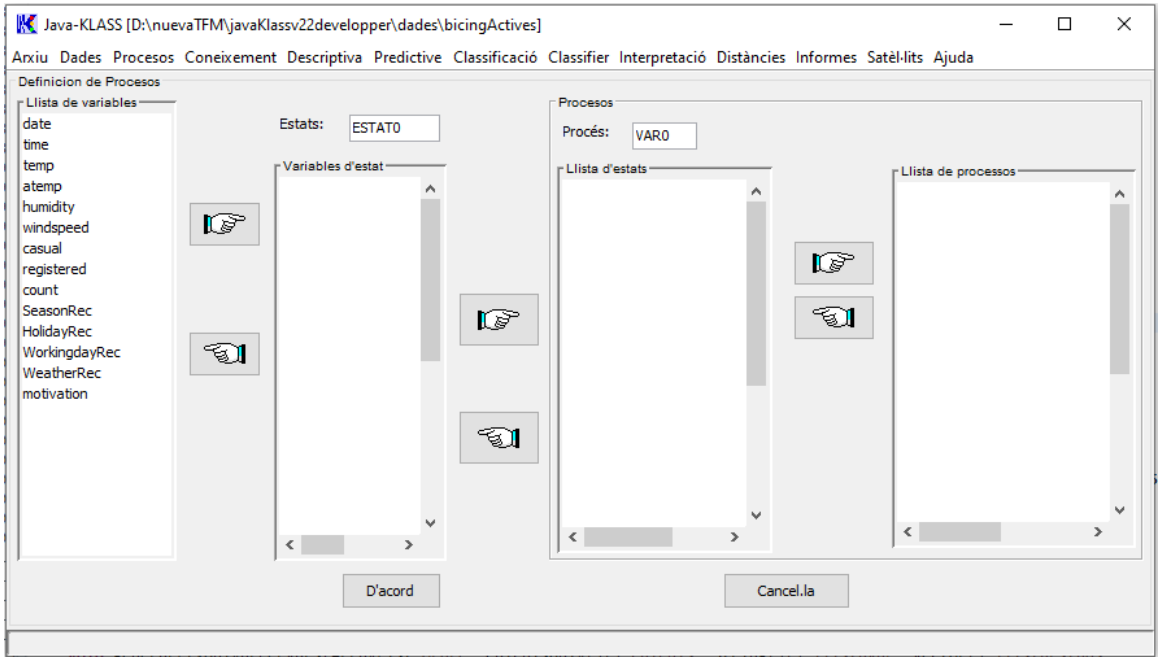


Figura 8 Interfaz definición de procesos

2.3.4.2 Mejora a la funcionalidad Clasificación por estados

La solución que se diseñó para mejorar esta clasificación fue:

Ampliar la estructura del Gestor Matriu con una propiedad nueva de la clase que es una lista de procesos asociados a la matriz de datos en el Gestor Matriu e intervenir en el código que recibe la creación de un nuevo proceso para que lo inserte en dicha lista.

El proceso en si sigue siendo un objeto con estructura de lista de estados asociados y los estados a su vez se asocian a listas de variables de las que están definidas en l_props (lista de propiedades, estructura de almacenamiento con la información relacionadas a las variables, para el manejo dentro de KLASS)

Se ha intervenido en todas las zonas donde KLASS consultaba los procesos para asegurar que trabaja con la lista entera de procesos y se puede elegir el que se quiere tratar o se han añadido parámetros para identificar el proceso donde se trata un proceso concreto

Por otro lado, se ha intervenido en la Interfaz de Clasificación por estados, (que recoge los parámetros de usuario de ejecución para una familia de clasificaciones por estados más general que la ClBRxE) para introducir la posibilidad que el usuario seleccione de una lista de procesos definidos sobre la matriz de datos aquel proceso con el cual se va a realizar la clasificación por estados, y se visualizará todos los estados que componen el proceso

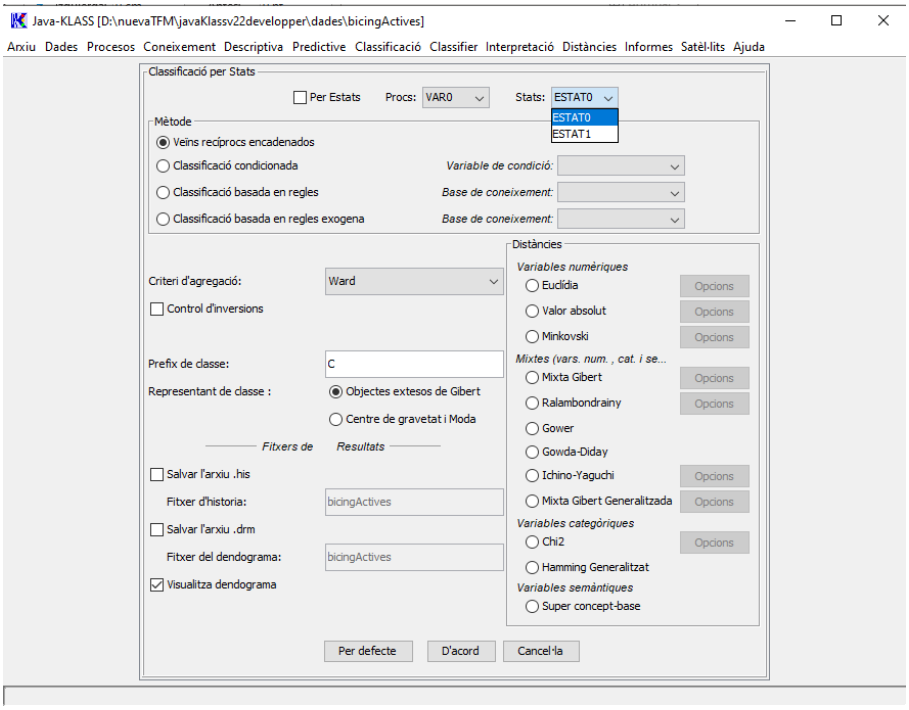


Figura 9 Interfaz Clasificación por estados

De acuerdo con la definición de la metodología de Clasificación por estados, KLASS realizará una clasificación local a cada uno de los estados utilizando las variables de dicho estado en cada ocasión y producirá una lista de dendrogramas, correspondientes a cada uno de los estados del proceso en análisis

Una vez obtenidos estos dendrogramas, el usuario los podrá visionar con las herramientas habituales que proporciona KLASS y decidir el nivel del corte. Al final de esta etapa, la matriz de datos contendrá E nuevas variables categóricas, una para estado, correspondientes a las clasificaciones resultantes de cortar estos dendrogramas

2.3.4.3 Implementación del análisis de trayectorias en datos de proceso discretos

En esta sección detallaremos todo el proceso para la implementación del análisis de trayectorias y la generación del diagrama de trayectorias de variables cualitativas, para lo cual se desarrollaron las siguientes clases:

Clase	TrajectoryGraph	
Descripción	TrajectoryGraph es la clase que sostendrá el gráfico de trayectorias. Contiene dos propiedades: • La lista de variables seleccionadas para intervenir en la trayectoria • El nombre de la variable resultante que contendrá las trayectorias y que eventualmente se guardará en la matriz de datos, a elección del usuario.	
Campos	Tipo de dato	Descripción
etiqueta	String	Etiqueta que contendrá el reporte y el nombre de la columna en el caso que el investigador decida conservar la variable resultante que contiene la trayectoria
lista_modalidad_trayectoria	ArrayList<ModalidadNodoTrayectoria>	Lista modalidades que contendrá cada uno de los nodos

Tabla 1 Clase trajectoryGraph

Clase	NodoTrayectoria	
Descripción	Clase que contiene información de una de las variables cualitativas “e” a incluir en el diagrama de trayectorias . Para esta investigación corresponde con una de las variables de estado del proceso a analizar	
Campos	Tipo de dato	Descripción
etiqueta	String	Es el identificador de la variable de estado a tratar e
lista_modalidad_trayectoria	ArrayList<ModalidadNodoTrayectoria>	Lista Nodos del diagrama de trayectorias correspondientes a la variable de estado e

Tabla 2 Clase NodoTrayectoria

Clase	ModalidadNodoTrayectoria	
Descripción	Contiene la información gráfica de un nodo del diagrama de trayectorias. En él está establecida la ubicación geométrica en el gráfico, la etiqueta del nodo y esta es la información que se utilizará para dibujar el nodo en LaTeX,	
Campos	Tipo de dato	Descripción
etiqueta	String	Etiqueta de un nodo del diagrama de trayectorias
x	Int	Posición x del nodo en el diagrama de trayectorias
Y	Int	Posición y del nodo en el diagrama de trayectorias

Tabla 3 ModalidadNodoTrayectoria

Clase	ModalidadFrecuencia	
Descripción	Contiene información sobre una trayectoria	
Campos	Tipo de dato	Descripción
modalidad	String	Secuencia de nodos por los que pasa la trayectoria (se separan con el signo "+")
Frecuencia	Float	Frecuencia de la trayectoria en los datos analizados (número de veces que aparece dicha trayectoria respecto al total de observaciones)

Tabla 4 ModalidadFrecuencia

A continuación, detallaremos los métodos creados durante la implementación de la Tesis:

UI (interficie de usuario de KLASS)

Panel Diagrama de Trayectoria
El diagrama de trayectorias es un componente desarrollado en esta tesis de master, con el objetivo de generar un gráfico de trayectorias a partir de los caminos generados de las variables cualitativas construidas por la clasificación basada en reglas de procesos, sin embargo, tolera cualquier variable cualitativa.

Los componentes que interactúan con este módulo son: GestorMatriu, GestorKlass, Colors y GeneratorLatex.

Validar Formulario
Verifica si el formulario tiene algún tipo de error que pueda generar inconsistencia al generar el gráfico de trayectoria

Validar número máximo de camino
Consulta el número máximo de caminos que se pueden generar y valida que no pueda exceder ese número y que sea mayor que 0

Construir diagrama
Genera un diagrama temporal para poder identificar el número máximo de caminos que tiene ese diagrama antes de generar el reporte

Ejecutar informe

Construye el reporte de LaTeX, realizando las respectivas validaciones y enviando a construir a partir de la estructura de Gráfico de trayectoria, los argumentos que envían para realizar este proceso son: Nombre de variable de trayectoria, número de trayectorias, conservar variable, ver informe, y gama normalizada

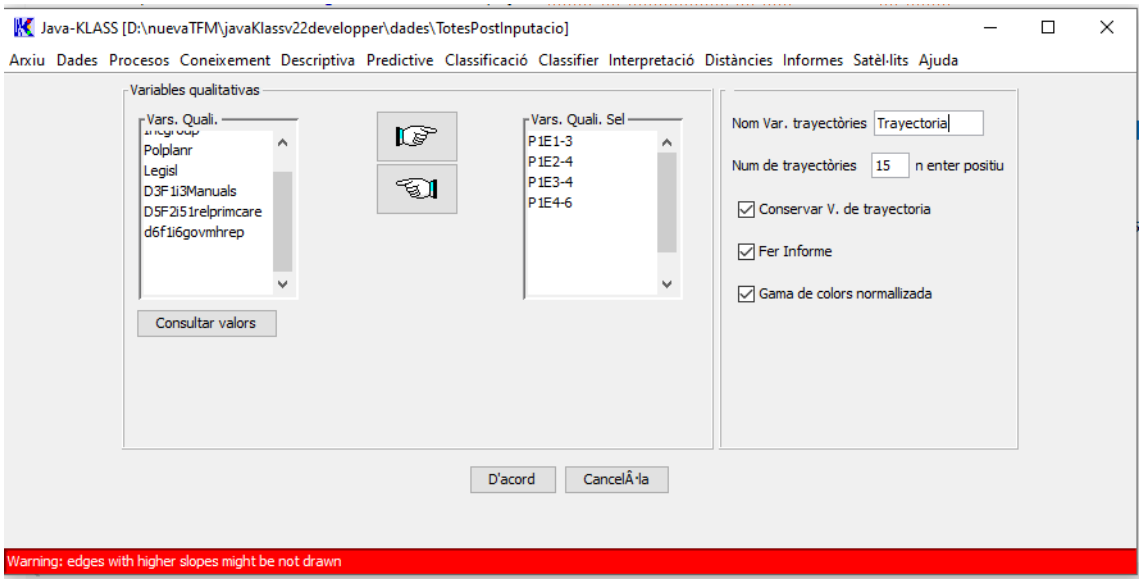


Figura 10 Panel diagrama trayectoria

NUCLI (núcleo de KLASS)

Gestor Klass

Construir diagrama

El instanciamiento del gráfico de trayectoria usuario trayectorias, se enviará la lista de variables cualitativas, y se escogerán las generadas a partir de haber podado el dendrograma generado por la clasificación basada en reglas de estados.

Se carga la lista de Nodos trayectorias a partir de la lista de que son representadas por las variables generadas por el dendrograma, a partir de su construcción se carga la lista de sus modalidades

Generar Diagrama de trayectorias

Permite instanciar el gráfico de trayectorias y e invocar a un método del gestor matriu para crear el diagrama de trayectorias.

Se calcula la posición espacial de cada uno de los Modalidad Nodo Trayectoria, para que se visualice en el diagrama de trayectoria.

Trayectory Graph

Obtener nodo trayectoria
Obtiene el nodo de una trayectoria por su nombre

Obtener número máximo de modalidades Graph
Determina el número máximo de modalidades que van a estar representadas

Obtener número máximo de modalidades a partir de una lista
A partir de una lista de elementos, se obtiene el número máximo de modalidades que representará en el gráfico, esto permite calcular la dimensión máxima y gestionarlo con el zoom

Obtener nombre
Obtener el nombre de la etiqueta del nodo, en este caso sería la etiqueta de la variable cualitativa generada por la clasificación basada en reglas de estados

Agregar nodo trayectoria
Agrega la variable cualitativa al gráfico.

Obtener Modalidad Nodo trayectoria
Retorna la modalidad de una trayectoria a partir del índice

Obtener índice de Modalidad Nodo trayectoria
Obtiene el índice de la lista de la modalidad de un nodo de la trayectoria

Obtener número máximo de caminos
Devuelve el número máximo de caminos que se pueden generar

Obtener número máximo de caminos por lista
Devuelve el número máximo de los caminos seleccionados por el usuario

Nodo Trayectoria

Agregar modalidad trayectoria
Agrega una lista de modalidades al nodo de la trayectoria

Obtener lista de modalidades del nodo trayectoria
Retorna la lista de modalidades del nodo de trayectoria

Obtener etiqueta
Retorna la etiqueta del Nodo de la trayectoria

Modalidad nodo trayectoria

Obtener etiqueta
Obtiene el nombre de la etiqueta correspondiente a la modalidad que estará representada en la gráfica de trayectoria.

Establecer posición
Cada nodo representado en la gráfica del reporte de gráfica de trayectoria tiene una ubicación espacial al ingresar la información se calcula la ubicación que debe de estar en el diagrama del reporte generado.

Modalidad Frecuencia
Contiene los atributos de etiqueta y probabilidad, consiste en almacenar la información del camino concatenado y su frecuencia, posteriormente para luego ser ordenado de mayor a menor

Obtener frecuencia
Devuelve la frecuencia que es construida por la tabla de frecuencia

Comparar
Permite comparar para realizar el ordenamiento de mayor a menor de la lista de Modalidades Frecuencia, gracias a esto se identificar los caminos más frecuentes

Tiene un trayecto anterior
Consulta si la tiene un camino anterior o si es el primer elemento del camino

2.3.4.4 Modelo de color asociado a las trayectorias

Las trayectorias entrañan dos aspectos gráficos: Por dónde pasan y qué tan importantes son

En el apartado anterior se ha detallado cómo ubicar las posiciones de los arcos del diagrama para representar las trayectorias

Sin embargo, en cuanto en un gráfico se cruzan dos trayectorias sobre el mismo espacio de nodos, es necesario introducir un distintivo gráfico entre ellas para poder identificar por donde discurre cada una. En este aspecto se puede trabajar con el grosor el trazo de la línea o con un modelo de coloreado automático

En este trabajo se ha optado por aprovechar el modelo de color diseñado para la construcción de los cuadros semáforos que ya tenía KLASS [Gibert, Conti 2015] y se ha adaptado para su utilización en el diagrama de trayectorias.

Los principios básicos que se han seguido son:

- las trayectorias se pintarán de colores proporcionales a su frecuencia
- Las trayectorias más frecuentes serán verdes, las más raras, rojas y las intermedias, tendrán un color graduado según el modelo citado entre los tonos rojo verde y amarillo
- El gráfico incluirá una leyenda que permitirá conocer la frecuencia de las trayectorias asociadas a cada color

  El modelo original presentado en [Gibert, Conti 2015] especificaba el color en base a un Tono más una frecuencia f que evaluaba en el intervalo real [0,1] y originalmente medía la variancia de una clase para la variable en estudio. El modelo de color en cuestión representaba una función c(t,f) que dado un tono y una frecuencia devolvía un color de la gama del tono t que iba del color puro (cuando f =0) a un color de la misma gama muy embrutecido y oscuro, tanto más cuanto más se acerca f a 1. El modelo calcula las ecuaciones que dan la gradación de color en tres tonos básicos (rojo, verde y amarillo)

  En esta tesis, se aprovecha c(t,f) para calcular el color con el que se pintará una trayectoria concreta. La integración del color en las trayectorias se realiza de la siguiente forma: f corresponderá a frecuencia relativa de una trayectoria concreta y t se maneja de la siguiente forma: t=rojo cuando f se encuentra en el rango de [0, 0.33], t= Amarillo para f entre (0.33,0.66] y t= verde si f >0.066

  Para dibujar la línea con el camino y su respectivo color, se toma la lista de Modalidades Frecuencia, el contenido de ella se procesa para obtener la secuencia de modalidades que sigue la trayectoria y en cada estado, se buscan las coordenadas de los nodos de origen y destino para trazar la línea desde la modalidad Mí, x a Mi+1, y, de esta manera se representa de forma visual los caminos en el diagrama de trayectorias.

  Además, el gráfico contiene una leyenda que permite interpretar las frecuencias de las trayectorias representadas

2.3.4.5 Construcción del gráfico de trayectorias

Generador del reporte

De acuerdo con la filosofía que gobierna el sistema KLASS la mayor parte de sus resultados se crean en código fuente de LaTeX en tiempo de ejecución y automáticamente se compilan y visualizan.

GeneradorText

Generador de imagen diagrama Trayectoria
Para generar el diagrama de trayectoria es necesario tener la lista de **Nodo Trayectorias**,

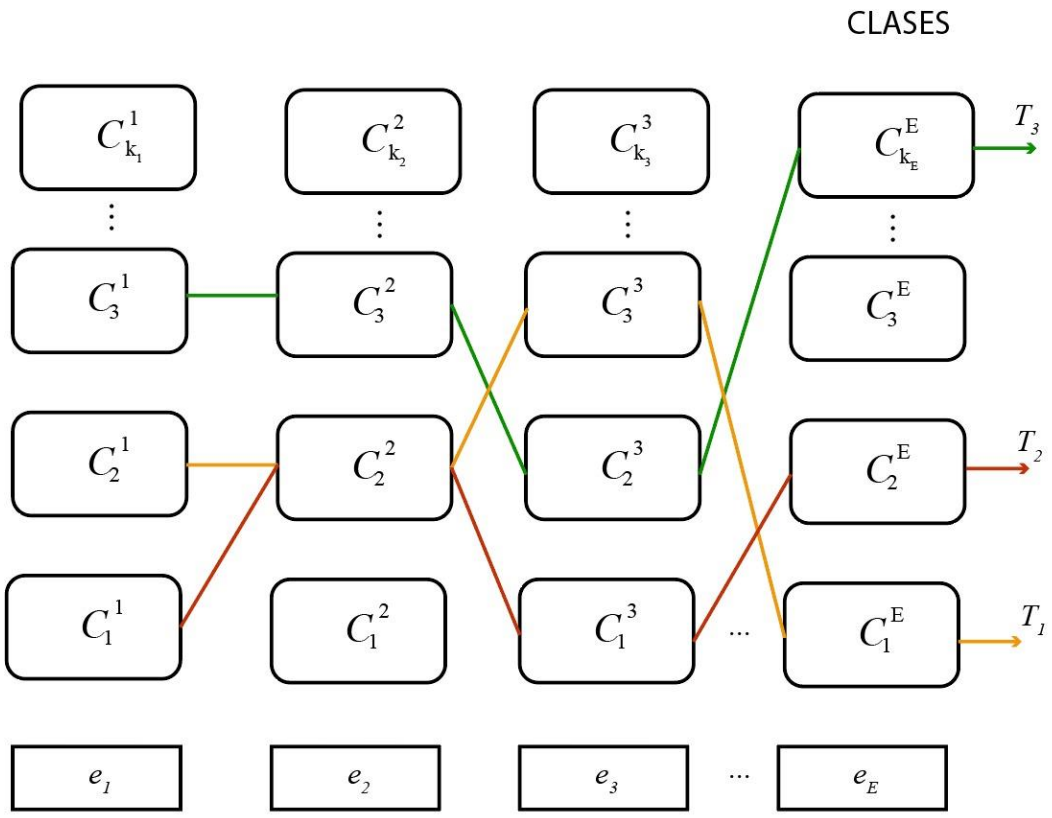


Figura 11 Esquema diagrama Trayectorias

El objeto Gráfico de trayectoria contiene una lista de Nodos de Trayectoria, NT_1, NT_2, NT_n, que se corresponden con la lista de estados que forman el proceso el cuál es representado en el gráfico en la parte inferior. Cada estado se asocia a su vez a una lista de clases o Modalidades desplegadas sobre el eje Y,

Si el gráfico contiene algún estado con más de 13 modalidades se aplicará un zoom delimitando el tamaño de la figura dentro de la página del reporte.

En el reporte de resultados sobre el proceso de construcción del diagrama de trayectorias

- Estructura del reporte
- Preámbulo
- Numero de trayectorias generadas
- Numero de trayectorias representada
- Lista de las variables de estado
- Títulos

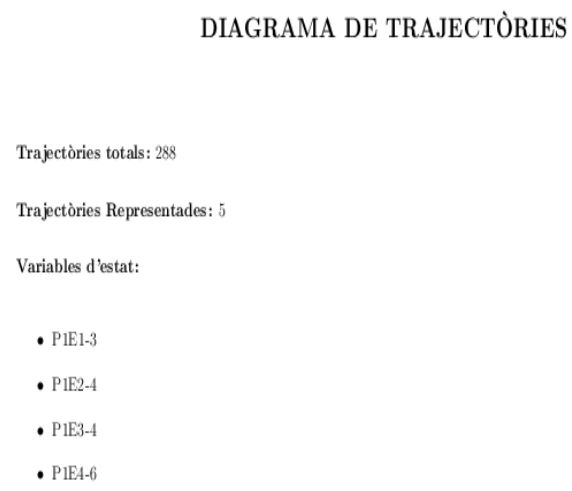


DIAGRAMA DE TRAJECTÒRIES

Trajectòries totals: 288

Trajectòries Representades: 5

Variables d'estat:

- P1E1-3
- P1E2-4
- P1E3-4
- P1E4-6

Figura 12 cabecera Diagrama trayectoria

- Diagrama de trayectoria con leyenda de colores de las frecuencias que se encuentra representando

- Reporte con el análisis descriptivo de la variable de trayectoria generada, y de las variables de clases que conforman la trayectoria generada, este parte del reporte contiene:

    - Tablas de frecuencia y diagramas de barras de la variable de trayectoria
    - Tabla de frecuencia y diagrama de barra de las variables que intervienen en la trayectoria

Así, se ha creado el método GenerarDiagramatrayectoria que se describe a continuación

GenerarDiagramatrayectoria

KLASS invoca a este método para generar el reporte en formato LaTeX, en el cual contiene una estructura preestablecida por el software en él que contiene un preámbulo y el contenido en este caso es la imagen del diagrama de trayectorias, para lo cual deberá invocar la función Generar Imagen Diagrama Trayectoria.

Que genera el código LaTeX de un diagrama de trayectorias a partir de un objeto de tipo TrajectoryGraph que contiene toda la información gráfica y cualitativa necesaria para construirlo

Se define un lienzo de dimensiones [420 x 650] puntos que será el punto de partida de nuestro diagrama. Como condiciones iniciales se considera el tamaño original del diagrama para realizar cálculos

La anchura del lienzo será dividida por el número de estados que comprenden el proceso y se reparte la separación entre las columnas de los nodos, en el ancho definido para el tamaño original

La altura del lienzo será dividida por el número de clases del estado y se reparte la separación vertical entre los nodos.

Esto determina las coordenadas originales de todos los nodos del diagrama de trayectorias. Los nodos se representan con óvalos de 60pts de ancho y 30 pts. de alto.

Los estados se etiquetan abajo. Las clases de un mismo estado se posicionan equidistantes de abajo a arriba en una misma columna.

Los nodos se etiquetan con el propio nombre de las clases

Además, se reserva un espacio a la izquierda de la gráfica para la leyenda con el código de colores de las trayectorias

Recordamos que una vez determinadas las posiciones de todos los nodos del gráfico se generan instrucciones de código LaTeX que construyen la figura en el entorno *picture.*

Esta representación puede superar el espacio gráfico de una página si el número de estados o de clases por cada estado es muy grande. Para resolver esto hay dos políticas. Una consiste en reescalar el gráfico a base de recalcular las coordenadas de cada nodo y las propias dimensiones del gráfico de acuerdo con el nivel de reducción necesario para

circunscribir el gráfico en una página. La otra consiste en trabajar con el concepto de reescalado del gráfico en su totalidad. Puesto que LaTeX dispone de herramientas de reescalado global que permiten que un punto (o cualquier unidad de medida utilizada) se represente en el documento compilado ocupando un espacio reescalado, se ha considerado que era más eficiente y legible mantener el algoritmo de generación del código fuente de la figura y añadir por delante un proceso de cálculo del nivel de reescalado de la figura en su totalidad.

Después de un análisis detallado se ha observado que los 400pts de ancho definidos para el eje X permiten la colocación razonable de hasta 5 estados (60pts de ancho + 20pts de separación entre estados). Para dibujar más cantidad de estados se aplicaría una reducción total del gráfico con un ratio de reducción que se calculamos de la siguiente forma

zoomX= 420 / ((60 + 20) * número Estados)

Análogamente se observa que si alguno de los estados del gráfico contiene más de 13 (30pts de altura y 20pts de separación) es también necesario reducir las dimensiones del gráfico. Para representar estados con un número mayor de clases es necesario aplicar un zoom al diagrama, que se calcula de la siguiente forma:

zoomY = 650 / ((30 + 20) * NumeroMaximoClases)

siendo el NumeroMaximoClases el número de clases del estado con más clases del proceso que se está representando

Así, para procesos de hasta 5 estados y de 13 clases como máximo en los estados se representará el diagrama original. Cuando incremente una de las dos dimensiones se calcularán los dos ratios de zoom indicados para el eje X o el eje Y y se utilizará el mayor de los dos, garantizando así que la figura cabe en una sola página.

La instrucción \setlength{\unitlength}{<Ratio> pt} de LaTeX Permite redefinir la unidad de medida original (puntos en este caso) por una reinterpretación reducida (o ampliada) que permite reescalar el gráfico entero sin necesidad de recalcular las coordenadas de sus elementos gráficos en el nuevo lienzo reducido.

2.3.4.6 Interfaz de usuario de generación de diagramas de trayectorias

Se añadió una nueva interfaz de usuario que permitirá la generación del reporte de diagrama de trayectorias, esta interfaz permitirá al experto opciones para la personalización del reporte, tales como:

- Lista de variables cualitativas a seleccionar
- Lista de variables cualitativas seleccionadas
- Nom Var. Trayectòria: opción para cambiar el nombre de la nueva variable creada para la trayectoria.
- Núm. de trayectòrias: representar n trayectorias de entre las trayectorias generadas
- Conservar v. de trayectoria: añadir la variable de trayectoria en la matriz de datos o no añadirla a la matriz de datos
- Fer informe: esta opción permite visualizar en el informe final, la descripta univariable de la nueva variable y de las variables que intervienen en el cálculo de trayectoria
- Gama de colors normalizada: nos permite seleccionar la modalidad para pintar los arcos de trayectorias del diagrama, permitiendo escoger entre dos opciones: Por frecuencia o Por frecuencia Normalizada

El panel diseñado para interactuar con el usuario es el siguiente:

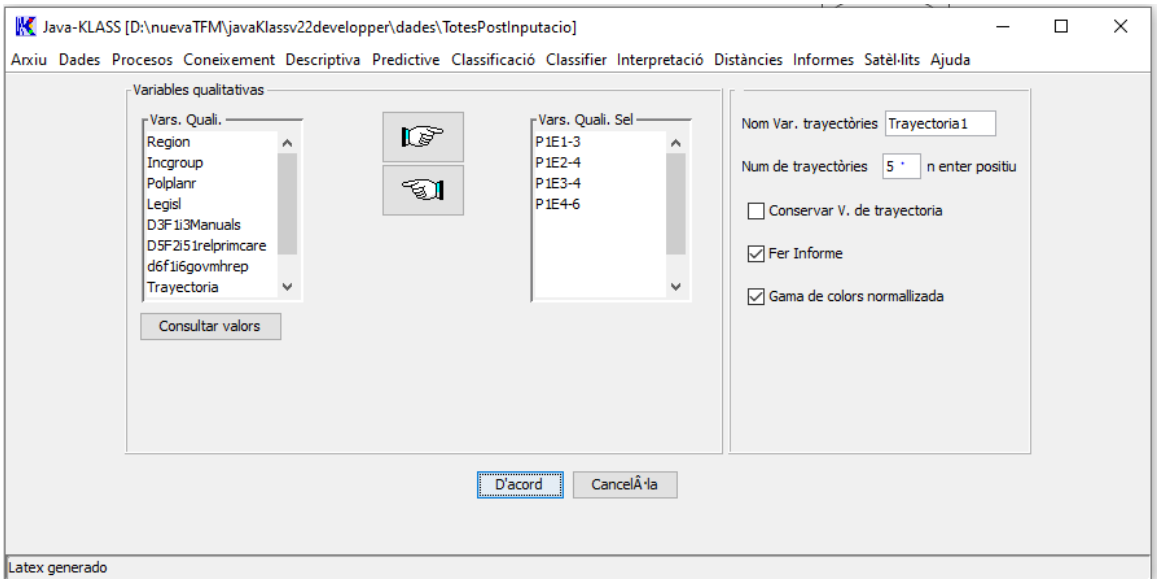


Figura 13 Panel diagrama de trayectòria

Una vez seleccionado los parámetros que requiere el experto, se genera un reporte compilado, ejecutado y visualizado en LaTeX en el que está representado un diagrama de trayectoria

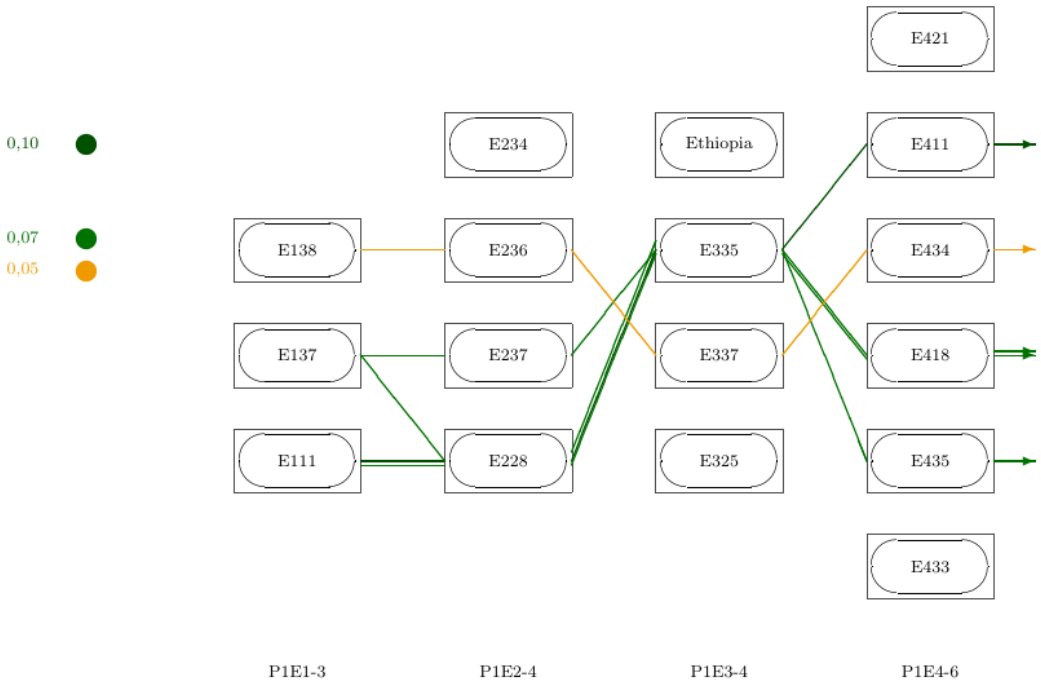


Figura 14 Ejemplo Diagrama de trayectoria generado

Capítulo 3

Aplicación a casos reales

En esta sección presentaremos algunos casos de estudio, definiendo los dataset utilizados y los procesos asociados para ilustrar el funcionamiento de la nueva funcionalidad añadida a KLASS en la presente tesis de máster

El esquema que seguiremos para esta experimentación será de forma general el siguiente:

- Carga del dataset a procesar
- Definición del proceso o procesos asociados a los datos, realizando la creación de estados con las variables disponibles en el dataset.
- Clasificación basada en reglas por estados: se realizará la clasificación basada en reglas por estado del proceso definido
- Análisis de la secuencia de dendrogramas generada. Se visualizará el dendrograma generado para cada estado y se realizará el corte de cada uno de ellos guardando las clasificaciones resultantes en la matriz.
- Generación de Diagrama de trayectoria, con las variables de clase creadas en el paso anterior (una para cada estado) y se generara el reporte con los resultados

A continuación, se procede al desarrollo de los casos, la presentación de los casos sometidos a la experimentación.

3.1 Caso de estudio Planta Depuradora de Agua Residuales

Para la presentación de este caso utilizaremos los datos de una planta depuradora ubicada en Catalunya, tomando como referencia la tesis de PHD [Rodríguez 2009] en el cual encontraremos el desarrollo del caso en mayor detalle.
Los datos a analizar en el presente caso son parte de la colaboración en proyecto de investigación de varias plantas Depuradoras, el Laboratori d'Enginyeria Química y ambiental de Girona y el Equipo de Ingeniería del conocimiento y Aprendizaje Automático del Departamento de LSI de la Universidad Politécnica de Cataluña, relacionado con la Dra. Karina Gibert Directora de esta Tesis de master.
Para el tratamiento de Aguas Residuales se debe de pasar por diferentes operaciones y procesos. Combinaciones de agentes físicos, físicos y biológicas son parte del diagrama de proceso de una estación depuradora [Rodríguez 2009]
Se muestra un esquema típico del proceso de tratamiento de aguas residuales, tomado de [Rodríguez 2009]

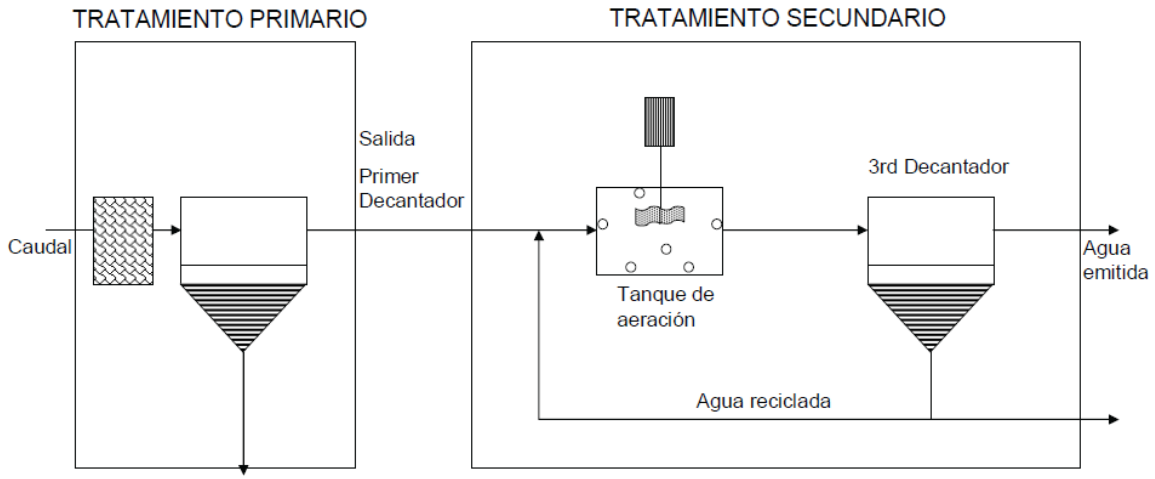


Figura 15 Esquema típico del proceso de tratamiento de aguas residuales.

3.1.1 Definición dataset

El dataset a trabajar está constituido por 396 observaciones obtenidos en el periodo de 1 año, con 40 variables tomadas de manera diarias, a continuación, detallaremos las 25 variables consideradas las más relevantes por los expertos y que serán parte de la clasificación definidas en 4 estados, que se usarán en este caso:

Variable	Descripción
Estado de Entrada - E	
Q-E	Caudal de entrada (metros cúbicos de agua por día)
FE-E	Pretratamiento con hierro (mg de hierro por litro de agua)
PH-E	pH (unidades de pH)
SS-E	Sólidos en suspensión (mg de sólidos por litro de agua)
SSV-E	Sólidos volátiles en suspensión (mg de sólidos por litro de agua)
DQO-E	Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)
DBO-E	Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)
Estado Decantador - D	
PH-D	pH (unidades de pH)
SS-D	Sólidos en suspensión (mg de sólidos por litro de agua)
SSV-D	Sólidos volátiles en suspensión (mg de sólidos por litro de agua)
DQO-D	Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)
DBO-D	Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)
Estado Bioreactor - B	
V30-B	Análisis volumétrico 30; calidad de sedimentación de la mezcla (ml por litro de agua)
MLSS-B	Sólidos en suspensión del licor mezcla (mg de sólidos por litro de licor mezcla)
MLVSS-B	Sólidos volátiles en suspensión del licor mezcla (mg de sólidos por litro de licor mezcla)
MCRT-B	Edad celular (días)
QB-B	Caudal del reactor biológico (metros cúbicos de agua por día)
QR-G	Caudal de recirculación (metros cúbicos de agua por día)
QP-G	Caudal de la purga (metros cúbicos de agua por día)
QA-G	Afluencia de aire (metros cúbicos de aire por día)
Estado Salida - S	
PH-S	pH (unidades de pH)
SS-S	Sólidos en suspensión (mg de sólidos por litro de agua)
SSV-S	Sólidos volátiles en suspensión (mg de sólidos por litro de agua)
DQO-S	Fracción de materia orgánica degradable por acción de agentes químicos oxidantes bajo condiciones de acidez (mg de oxígeno por litro de agua)

DBO-S	Fracción de materia orgánica biodegradable en agua residual (mg de oxígeno por litro de agua)

Tabla 5 Definición dataset Planta

3.1.2 Definición de procesos

Para definir un Proceso lo primero que se debe de realizar es la creación de los estados, las medidas corresponden a 4 estados o subprocesos según definición de la planta, estos son:

Estado Entrada denominado E con 7 variables: Q-E, FE-E, PH-E, SS-E, SSV-E, DQO-E, DBO-E

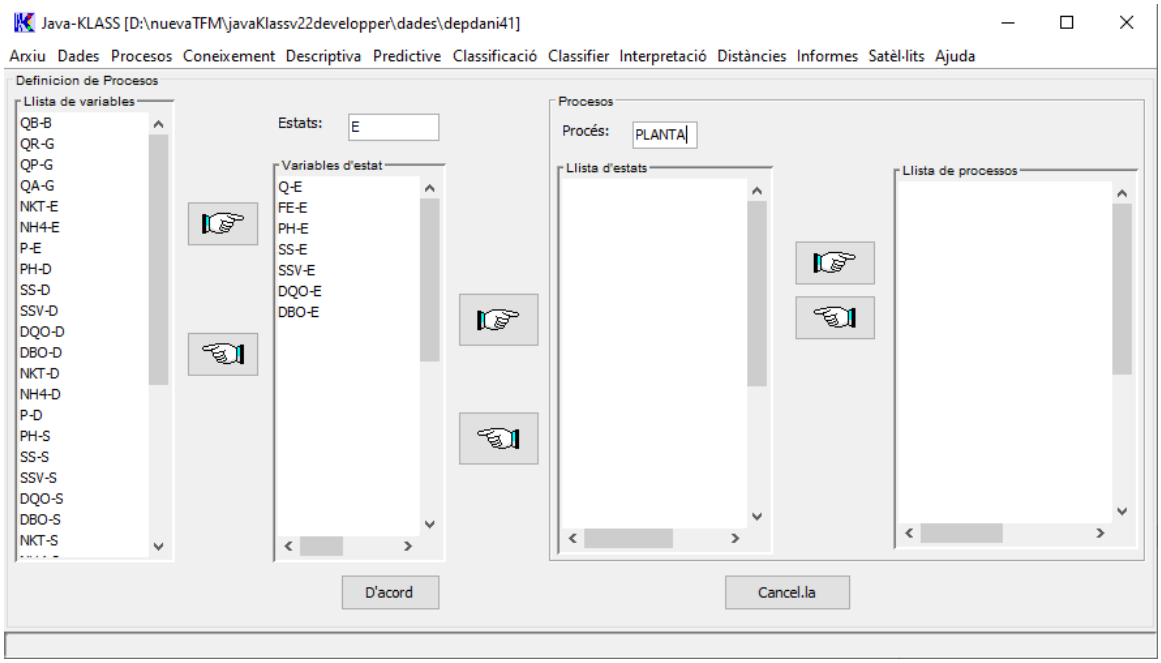


Figura 16 Estado Entrada - E

Estado Decantador denominado D con 5 variables: PH-D, SS-D, SSV-D, DQO-D, DBO-D

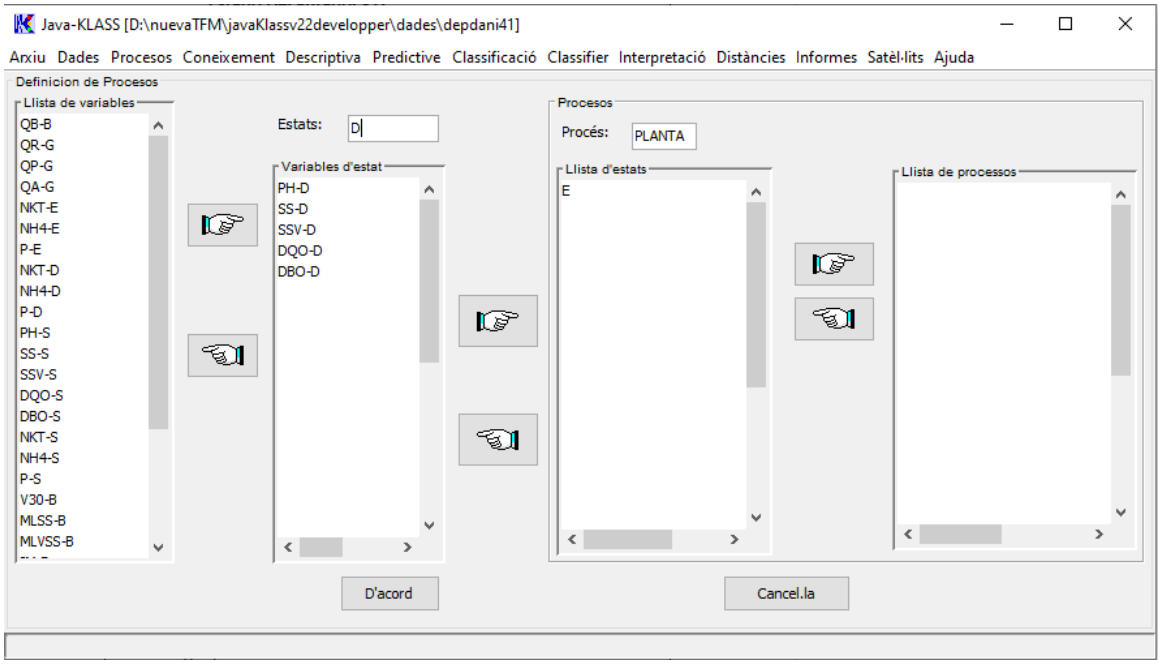


Figura 17 Estado Decantador - D

Estado Bioreactor denominado B con 8 variables: V30-B, MLSS-B, MLVSS-B, MCRT-B, QB-B, QR-G, QP-G, QA-G

Figura 18 Estado Biorreactor - B

Estado Salida denominado S con 5 variables: PH-S, SS-S, SSV-S, DQO-S, DBO-S

Figura 19 Estado Salida – S

Proceso Planta: con los estados E, D, B, S

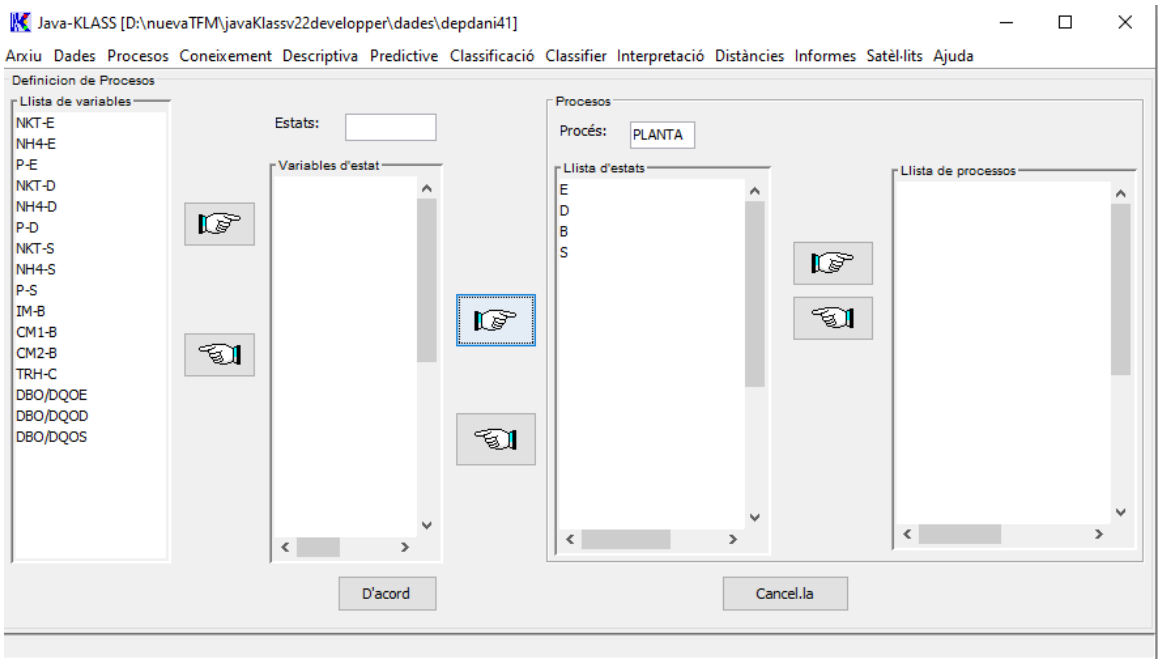


Figura 20 Definición Proceso Planta

3.1.3 Clasificaciones basadas en reglas por estados

Realizaremos la clasificación por estados del proceso Planta con los siguientes criterios: métrica euclidiana, agregación de Ward, como se muestra en la figura a continuación.

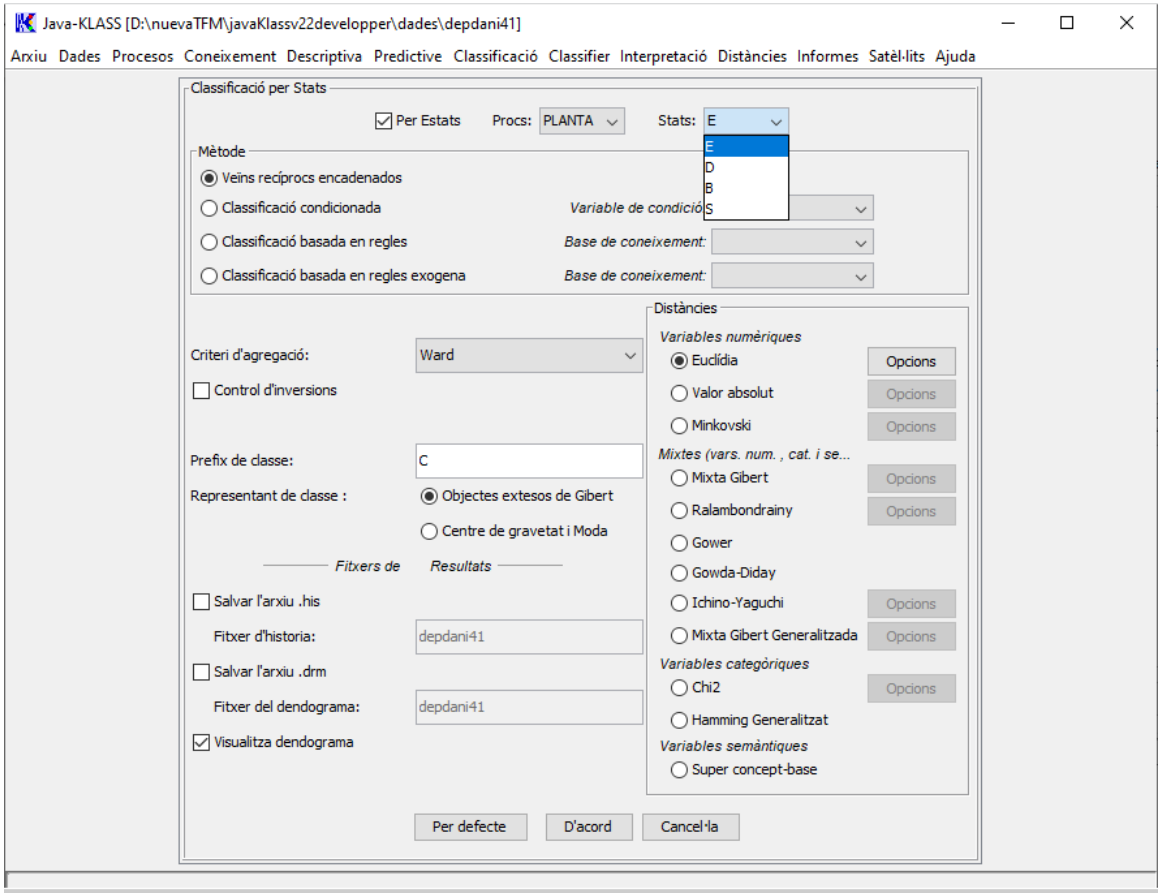


Figura 21 Clasificación por estados proceso Planta

La clasificación generará un dendrograma por cada estado clasificado, el cual deberá ser podado para maximizar su ratio, y así generar las variables de clases para su interpretación

A continuación, analizaremos cada una de la clase creada, el análisis descriptivo por clase completo se lo puede ver en el Anexo

Estado Entrada – E: el dendrograma generado por la clasificación para este estado, se lo ha cortado en 4 clases

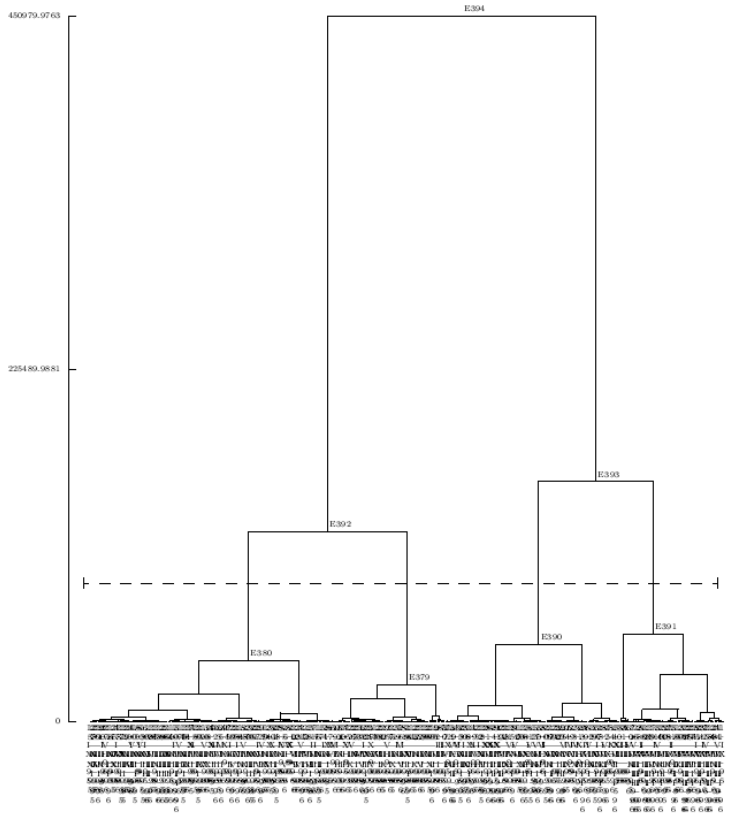


Figura 22 Corte Dendrogarma estado Entrada

A continuación, presentamos parte del análisis descriptivo del estado Entrada, para mayor información se puede ver el Anexo 1

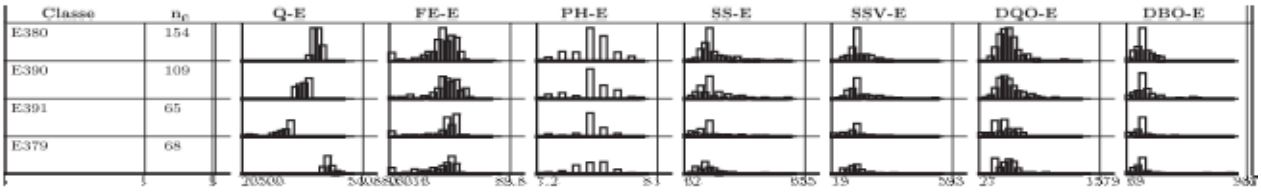


Figura 23 Panel de clases estado Entrada

A grandes trazos esta clasificación está dominada por el caudal de entrada a la planta. La clase E379 es la que mayor caudal ingresa, seguida de E380, E391 y la que menos es E390. Así, se proponen las siguientes etiquetas para las clases:

E379	CaudalPP
E380	CaudalP
E391	CaudalL
E390	CauldalLL

Tabla 6 Conceptualización de las clases de la Entrada de la planta

Estado decantador – D: el dendrograma generado por la clasificación para este estado, se lo ha cortado en 3 clases

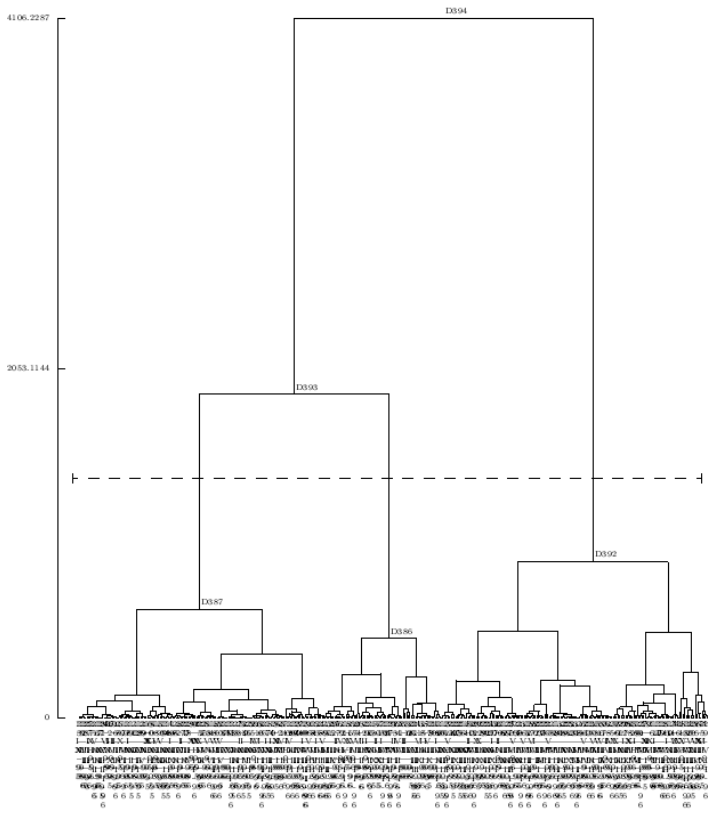


Figura 24 Corte Dendrograma estado Decantador

A continuación, presentamos parte del análisis descriptivo del estado Decantador, para mayor información se puede ver el Anexo 2

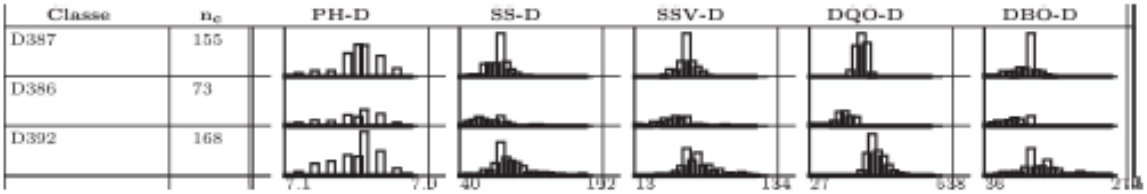


Figura 25 Panel de clases Estado Decantador

Utilizando la metodología habitual, se han identificado las variables significativas en las clases [Gibert, Sevilla-Villanueva, Sànchez-Marrè 2016] y con ayuda de los expertos se ha procedido al etiquetado de las mismas

	Etiqueta	PH-D	SS-D	SSV-D	DQO-D	DBO-D
D387	CargaMedia		MEDIO	MEDIO	MEDIO	BAJO
D386	Dilución		BAJO	BAJO	BAJO	BAJO
D392	CargaP		MEDIO	ALTO	MEDIO	ALTO

Tabla 7 Conceptualización de las clases Decantador de la planta

Estado Biorreactor – B: el dendrograma generado por la clasificación para este estado, se lo ha cortado en 3 clases

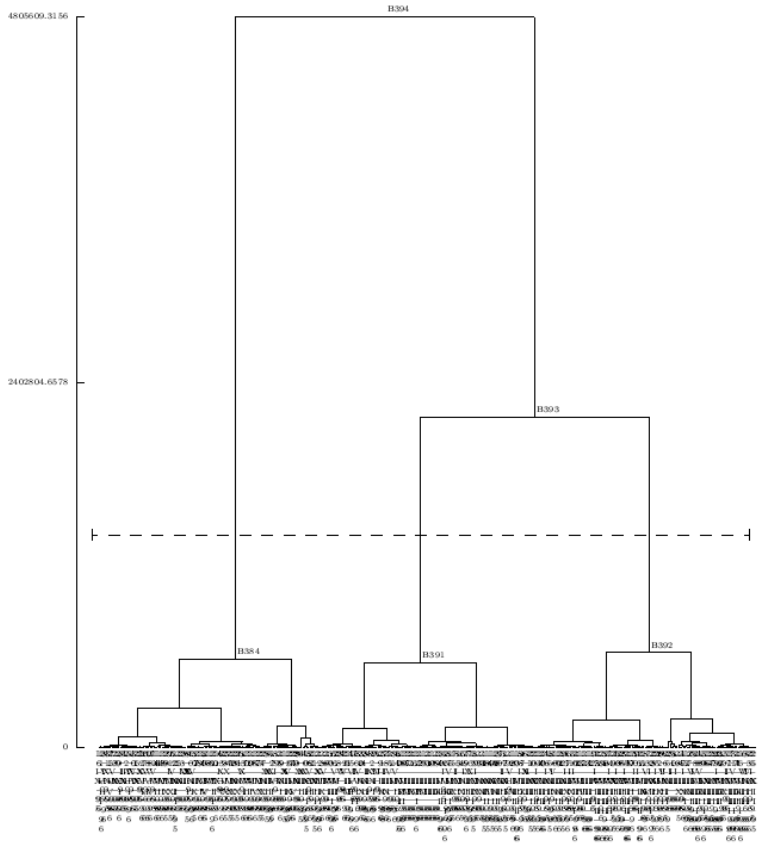


Figura 26Corte Dendrograma Estado Bioreactor

A continuación, presentamos parte del análisis descriptivo del estado Bioreactor, para mayor información se puede ver el Anexo 3

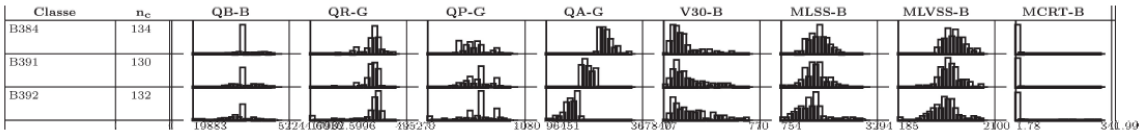


Figura 27 Panel de clases estado Bioreactor

En este caso es la aireación del tanque (QA-G) la que parece dominar la clasificación y la menos aireación se asocia con mayor V30 y menor tiempo de residencia de las células del reactor biológico (MLSS-B). Se proponen pues las siguientes etiquetas

B384	LodoActivo
B391	Intermedia
B392	LodoEnvejecido

Tabla 8 Conceptualización de las clases Bioreactor de la planta

Estado Salida – S: el dendrograma generado por la clasificación para este estado, se lo ha cortado en 3 clases

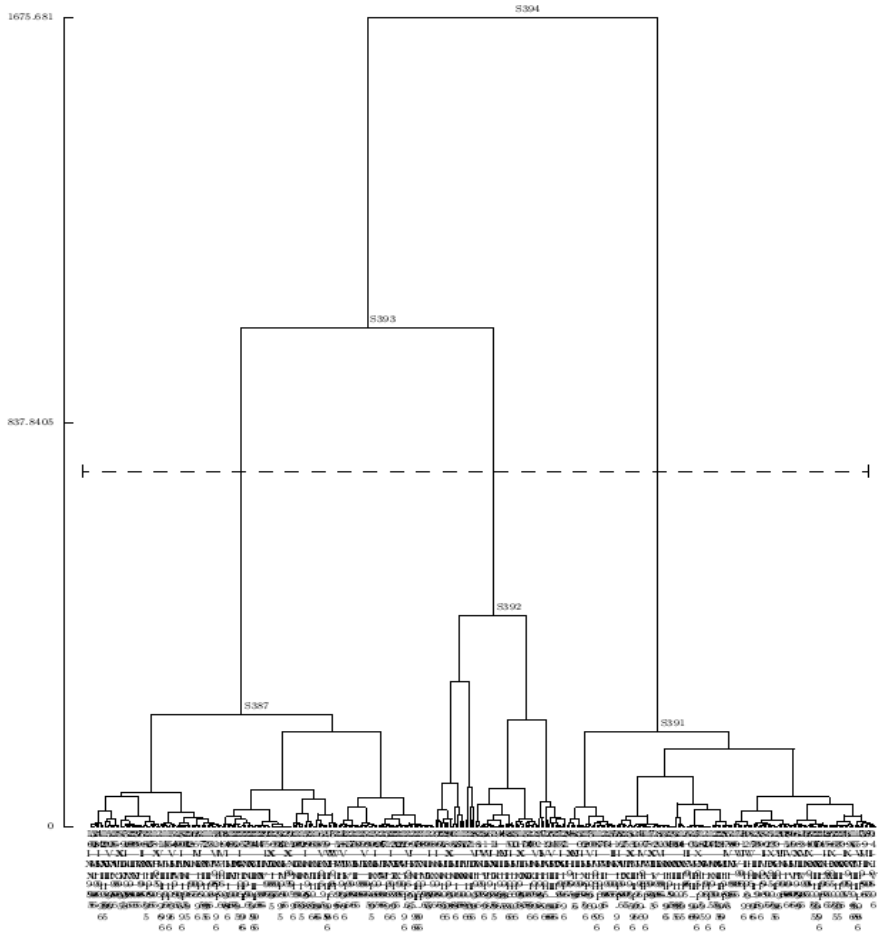


Figura 28 Corte Dendrogarma Estado Salida

A continuación, presentamos parte del análisis descriptivo del estado Salida, para mayor información se puede ver el Anexo 4

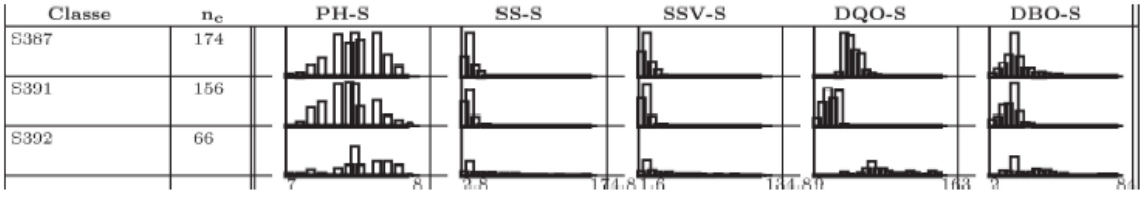


Figura 29 Panel de clases estado Salida

	Etiquetas	PH-S	SS-S	SSV-S	DQO-S	DBO-S
S387	conDQO				MEDIO	BAJO
S391	Limpia				BAJO	BAJO
S392	Carga				ALTO	ALTO

Tabla 9 Conceptualización de las clases de la salida de la planta

3.1.4 Generación de diagrama de trayectoria

Luego de haber realizado la clasificación por estados, procedemos a construir el diagrama de trayectorias

Eventualmente el experto podría introducir un paso intermedio de interpretación de las clases de cada estado a partir de los paneles de clases y proceder a un etiquetado de las clases, generar la recodificación de clases correspondiente y utilizar las variables recodificadas para construir el diagrama de trayectorias

Se va a representar las 5 trayectorias con frecuencia mayores, a partir de la tabla de frecuencia de la variable de trayectoria creada a partir de los 4 estados definidos en el proceso.

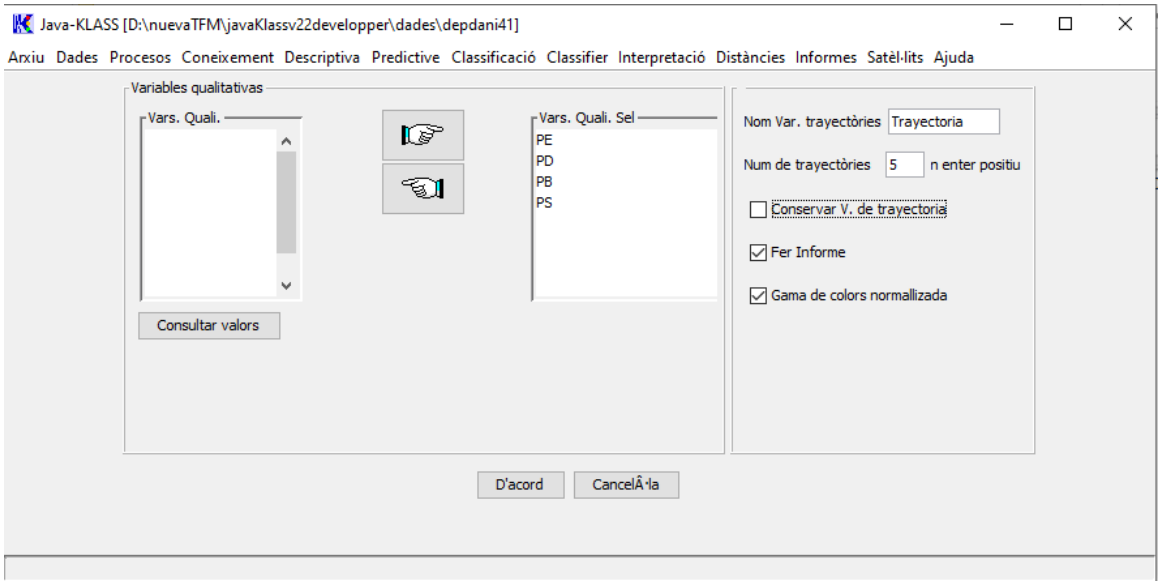


Figura 30 Panel diagrama Trayectoria proceso Planta

Tabla de frecuencias de la variable de trayectoria

Taula de freqüències				
Modalitats	Freq. absol.	Freq. acum.	Freq. relat.	Freq. rel. acum.
E380+D387+B384+S387	9	9	0.0227	0.0227
E390+D387+B384+S391	7	16	0.0177	0.0404
E391+D386+B391+S391	4	20	0.0101	0.0505
E380+D387+B384+S391	14	34	0.0354	0.0859
E379+D387+B384+S391	6	40	0.0152	0.101
E380+D387+B391+S387	9	49	0.0227	0.1237
E380+D386+B391+S391	5	54	0.0126	0.1364
E380+D392+B391+S392	7	61	0.0177	0.154
E390+D386+B392+S387	3	64	0.0076	0.1616
E391+D386+B391+S387	2	66	0.0051	0.1667
E390+D386+B392+S392	1	67	0.0025	0.1692
E390+D387+B391+S387	9	76	0.0227	0.1919
E380+D386+B392+S391	4	80	0.0101	0.202
E379+D386+B391+S391	6	86	0.0152	0.2172
E390+D386+B391+S392	1	87	0.0025	0.2197
E390+D387+B392+S392	3	90	0.0076	0.2273
E379+D387+B391+S387	4	94	0.0101	0.2374
E379+D387+B391+S391	6	100	0.0152	0.2525
E379+D392+B384+S392	1	101	0.0025	0.2551
E379+D392+B384+S387	11	112	0.0278	0.2828
E380+D392+B384+S387	19	131	0.048	0.3308
E380+D392+B384+S392	8	139	0.0202	0.351
E380+D392+B384+S391	7	146	0.0177	0.3687
E390+D392+B384+S392	3	149	0.0076	0.3763
E390+D386+B384+S387	5	154	0.0126	0.3889
E379+D392+B384+S391	5	159	0.0126	0.4015
E390+D386+B391+S391	2	161	0.0051	0.4066
E379+D387+B391+S392	2	163	0.0051	0.4116
E380+D386+B384+S387	2	165	0.0051	0.4167
E379+D387+B384+S392	1	166	0.0025	0.4192
E380+D387+B384+S392	2	168	0.0051	0.4242
E390+D386+B391+S387	4	172	0.0101	0.4343
E379+D387+B384+S387	7	179	0.0177	0.452
E390+D386+B384+S391	2	181	0.0051	0.4571
E390+D387+B384+S392	4	185	0.0101	0.4672
E390+D387+B384+S387	7	192	0.0177	0.4848
E390+D392+B384+S387	6	198	0.0152	0.5
E390+D392+B391+S391	5	203	0.0126	0.5126
E380+D392+B391+S391	6	209	0.0152	0.5278
E380+D387+B391+S391	16	225	0.0404	0.5682
E380+D392+B392+S387	13	238	0.0328	0.601

Tabla 10 Tabla de Frecuencias variable de trayectoria parte 1

Taul				
Modalit	Freq.	Freq.	Freq	F
E379+D386+B391	2	2	0.0051	0.
+S387	1	8	0.0025	7
E379+D392+B391	13	5	0.0328	1
+S387	3	2	0.0076	9
E391+D386+B392	1	8	0.0025	7
+S391	1	6	0.0025	0.
E391+D386+B392	3	2	0.0076	7
+S387	1	9	0.0025	2
E380+D386+B392	1	9	0.0025	2
+S387	3	3	0.0076	2
E379+D386+B392	1	0	0.0025	0.
+S387	2	2	0.0051	7
E379+D387+B392	2	3	0.0051	5
+S387	6	0	0.0152	5
E380+D386+B391	7	3	0.0177	1
+S387	3	3	0.0076	0.
E379+D392+B391	2	0	0.0051	7
+S392	2	4	0.0051	6
E391+D387+B391	9	3	0.0227	2
+S387	2	0	0.0051	6
E379+D386+B392	6	7	0.0152	0.
+S392	5	3	0.0126	7
E380+D387+B392	2	0	0.0051	6
		8		5
dades	0	N =	0	

Tabla 11 Tabla de Frecuencias variable de trayectoria parte2

Podemos observar en la variable de trayectoria que se han encontrado 83 trayectorias posibles cuyas frecuencias se representan en la siguiente figura

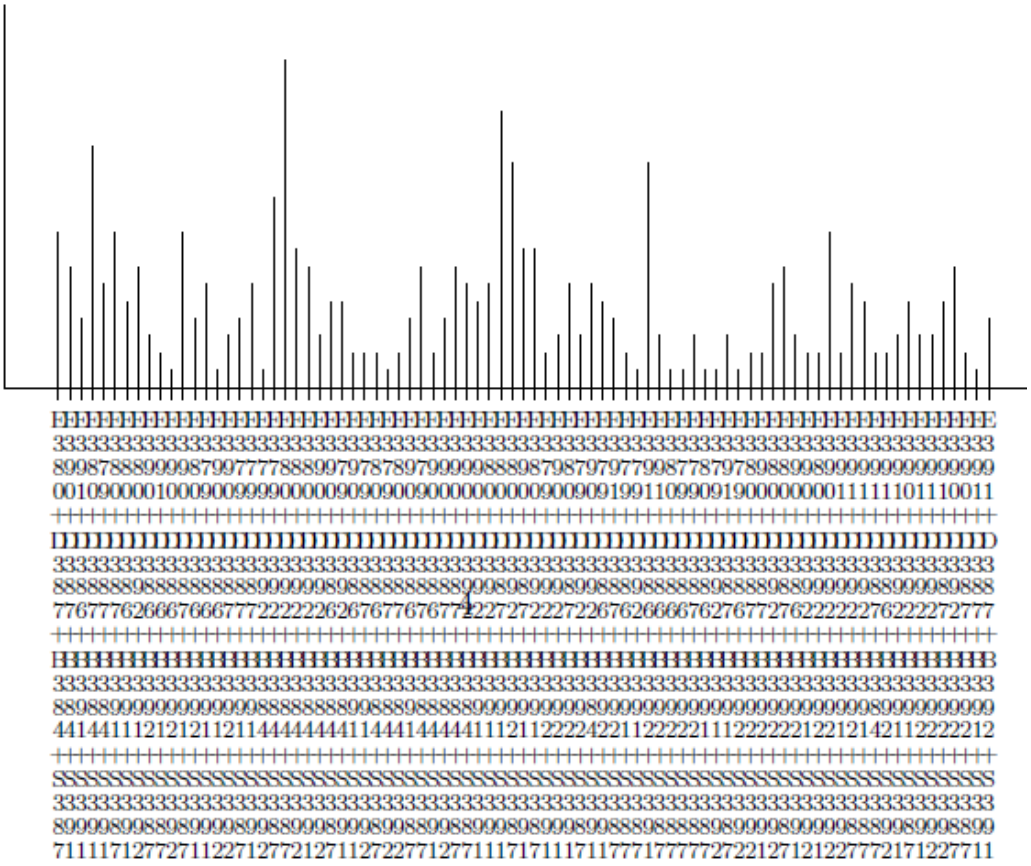


Figura 31 Diagrama barra Planta

Se han representado las 5 más frecuentes. Observamos que la trayectoria más frecuente es la E380+D392+B384+S387, la cual la podemos comparar en el diagrama de Barra y en el diagrama de trayectorias. Las trayectorias E380+D387+B391+S391 y E380+D387+B384+S391 son las 2 trayectorias que comparten la frecuencia de 0.04. Las trayectorias E380+D392+B392+S387, E391+D386+B392+S391 son con frecuencia 0.03, siguen en el orden de frecuencia

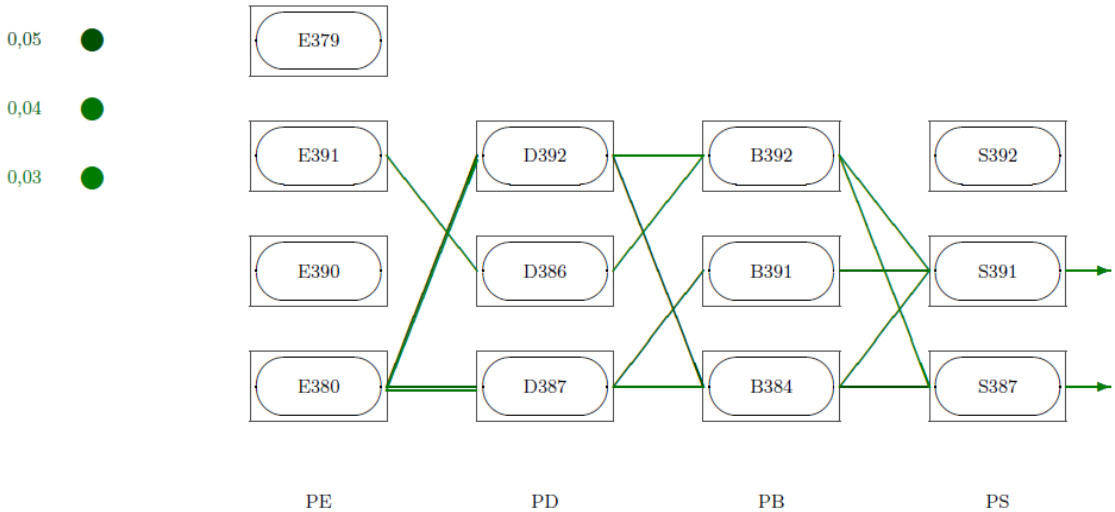


Figura 32 Diagrama de Trayectorias Proceso Planta 5 trayectorias más frecuentes

Para mayor información presentamos en el Anexo 5 el reporte completo del diagrama de trayectorias del proceso Planta

Si este diagrama se realiza con las clases etiquetadas por los expertos su potencia expresiva es mucho mayor

Variable Trayectoria

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
CaudalP+cargaMedia+LodoActivo+ConDQO	9	9	0.0227	0.0227
CaudalLL+cargaMedia+LodoActivo+Limpia	7	16	0.0177	0.0404
CaudalL+Dilucin+Intermedia+Limpia	4	20	0.0101	0.0505
CaudalP+cargaMedia+LodoActivo+Limpia	14	34	0.0354	0.0859
CaudalPP+cargaMedia+LodoActivo+Limpia	6	40	0.0152	0.101
CaudalP+cargaMedia+Intermedia+ConDQO	9	49	0.0227	0.1237
CaudalP+Dilucin+Intermedia+Limpia	5	54	0.0126	0.1364
CaudalP+CargaP+Intermedia+Carga	7	61	0.0177	0.154
CaudalLL+Dilucin+LodoEnvejecido+ConDQO	3	64	0.0076	0.1616
CaudalL+Dilucin+Intermedia+ConDQO	2	66	0.0051	0.1667
CaudalLL+Dilucin+LodoEnvejecido+Carga	1	67	0.0025	0.1692
CaudalLL+cargaMedia+Intermedia+ConDQO	9	76	0.0227	0.1919
CaudalP+Dilucin+LodoEnvejecido+Limpia	4	80	0.0101	0.202
CaudalPP+Dilucin+Intermedia+Limpia	6	86	0.0152	0.2172
CaudalLL+Dilucin+Intermedia+Carga	1	87	0.0025	0.2197
CaudalLL+cargaMedia+LodoEnvejecido+Carga	3	90	0.0076	0.2273
CaudalPP+cargaMedia+Intermedia+ConDQO	4	94	0.0101	0.2374
CaudalPP+cargaMedia+Intermedia+Limpia	6	100	0.0152	0.2525
CaudalPP+CargaP+LodoActivo+Carga	1	101	0.0025	0.2551
CaudalPP+CargaP+LodoActivo+ConDQO	11	112	0.0278	0.2828
CaudalP+CargaP+LodoActivo+ConDQO	19	131	0.048	0.3308
CaudalP+CargaP+LodoActivo+Carga	8	139	0.0202	0.351
CaudalP+CargaP+LodoActivo+Limpia	7	146	0.0177	0.3687
CaudalLL+CargaP+LodoActivo+Carga	3	149	0.0076	0.3763
CaudalLL+Dilucin+LodoActivo+ConDQO	5	154	0.0126	0.3889
CaudalPP+CargaP+LodoActivo+Limpia	5	159	0.0126	0.4015
CaudalLL+Dilucin+Intermedia+Limpia	2	161	0.0051	0.4066
CaudalPP+cargaMedia+Intermedia+Carga	2	163	0.0051	0.4116
CaudalP+Dilucin+LodoActivo+ConDQO	2	165	0.0051	0.4167
CaudalPP+cargaMedia+LodoActivo+Carga	1	166	0.0025	0.4192
CaudalP+cargaMedia+LodoActivo+Carga	2	168	0.0051	0.4242
CaudalLL+Dilucin+Intermedia+ConDQO	4	172	0.0101	0.4343
CaudalPP+cargaMedia+LodoActivo+ConDQO	7	179	0.0177	0.452
CaudalLL+Dilucin+LodoActivo+Limpia	2	181	0.0051	0.4571
CaudalLL+cargaMedia+LodoActivo+Carga	4	185	0.0101	0.4672
CaudalLL+cargaMedia+LodoActivo+ConDQO	7	192	0.0177	0.4848
CaudalLL+CargaP+LodoActivo+ConDQO	6	198	0.0152	0.5
CaudalLL+CargaP+Intermedia+Limpia	5	203	0.0126	0.5126
CaudalP+CargaP+Intermedia+Limpia	6	209	0.0152	0.5278
CaudalP+cargaMedia+Intermedia+Limpia	16	225	0.0404	0.5682
CaudalP+CargaP+LodoEnvejecido+ConDQO	13	238	0.0328	0.601
CaudalLL+cargaMedia+Intermedia+Limpia	8	246	0.0202	0.6212
CaudalP+CargaP+Intermedia+ConDQO	8	254	0.0202	0.6414
CaudalPP+CargaP+LodoEnvejecido+Limpia	2	256	0.0051	0.6465
CaudalLL+CargaP+LodoEnvejecido+Limpia	3	259	0.0076	0.654
CaudalP+cargaMedia+LodoEnvejecido+Limpia	6	265	0.0152	0.6692
CaudalPP+CargaP+LodoEnvejecido+ConDQO	3	268	0.0076	0.6768
CaudalLL+CargaP+LodoActivo+Limpia	6	274	0.0152	0.6919
CaudalPP+Dilucin+LodoEnvejecido+Limpia	5	279	0.0126	0.7045
CaudalL+cargaMedia+LodoEnvejecido+ConDQO	4	283	0.0101	0.7146

Tabla 12 Tabla de Frecuencias variable de trayectoria con Etiquetas parte 1

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
CaudalPP+Dilucin+Intermedia+ConDQO	2	285	0.0051	0.7197
CaudalPP+CargaP+Intermedia+ConDQO	1	286	0.0025	0.7222
CaudalL+Dilucin+LodoEnvejecido+Limpia	13	299	0.0328	0.7551
CaudalL+Dilucin+LodoEnvejecido+ConDQO	3	302	0.0076	0.7626
CaudalP+Dilucin+LodoEnvejecido+ConDQO	1	303	0.0025	0.7652
CaudalPP+Dilucin+LodoEnvejecido+ConDQO	1	304	0.0025	0.7677
CaudalPP+cargaMedia+LodoEnvejecido+ConDQO	3	307	0.0076	0.7753
CaudalP+Dilucin+Intermedia+ConDQO	1	308	0.0025	0.7778
CaudalPP+CargaP+Intermedia+Carga	1	309	0.0025	0.7803
CaudalL+cargaMedia+Intermedia+ConDQO	3	312	0.0076	0.7879
CaudalPP+Dilucin+LodoEnvejecido+Carga	1	313	0.0025	0.7904
CaudalP+cargaMedia+LodoEnvejecido+Carga	2	315	0.0051	0.7955
CaudalLL+cargaMedia+LodoEnvejecido+Limpia	2	317	0.0051	0.8005
CaudalP+CargaP+LodoEnvejecido+Carga	6	323	0.0152	0.8157
CaudalP+cargaMedia+LodoEnvejecido+ConDQO	7	330	0.0177	0.8333
CaudalLL+Dilucin+LodoEnvejecido+Limpia	3	333	0.0076	0.8409
CaudalLL+CargaP+Intermedia+Carga	2	335	0.0051	0.846
CaudalP+CargaP+LodoEnvejecido+Limpia	2	337	0.0051	0.851
CaudalLL+CargaP+LodoEnvejecido+Carga	9	346	0.0227	0.8737
CaudalL+CargaP+Intermedia+Carga	2	348	0.0051	0.8788
CaudalL+CargaP+LodoEnvejecido+ConDQO	6	354	0.0152	0.8939
CaudalL+CargaP+Intermedia+ConDQO	5	359	0.0126	0.9066
CaudalL+cargaMedia+LodoActivo+ConDQO	2	361	0.0051	0.9116
CaudalL+Dilucin+LodoEnvejecido+Carga	2	363	0.0051	0.9167
CaudalL+CargaP+Intermedia+Limpia	3	366	0.0076	0.9242
CaudalLL+CargaP+Intermedia+ConDQO	5	371	0.0126	0.9369
CaudalL+CargaP+LodoEnvejecido+Limpia	3	374	0.0076	0.9444
CaudalL+CargaP+LodoEnvejecido+Carga	3	377	0.0076	0.952
CaudalL+cargaMedia+LodoEnvejecido+Carga	5	382	0.0126	0.9646
CaudalLL+CargaP+LodoEnvejecido+ConDQO	7	389	0.0177	0.9823
CaudalLL+cargaMedia+LodoEnvejecido+ConDQO	2	391	0.0051	0.9874
CaudalL+cargaMedia+Intermedia+Limpia	1	392	0.0025	0.9899
CaudalL+cargaMedia+LodoEnvejecido+Limpia	4	396	0.0101	1
dades mancants	0	N = 396	0	

Tabla 13 Tabla de Frecuencias variable de trayectoria con Etiquetas parte 2

Se presenta el diagrama de trayectorias con las clases etiquetadas.

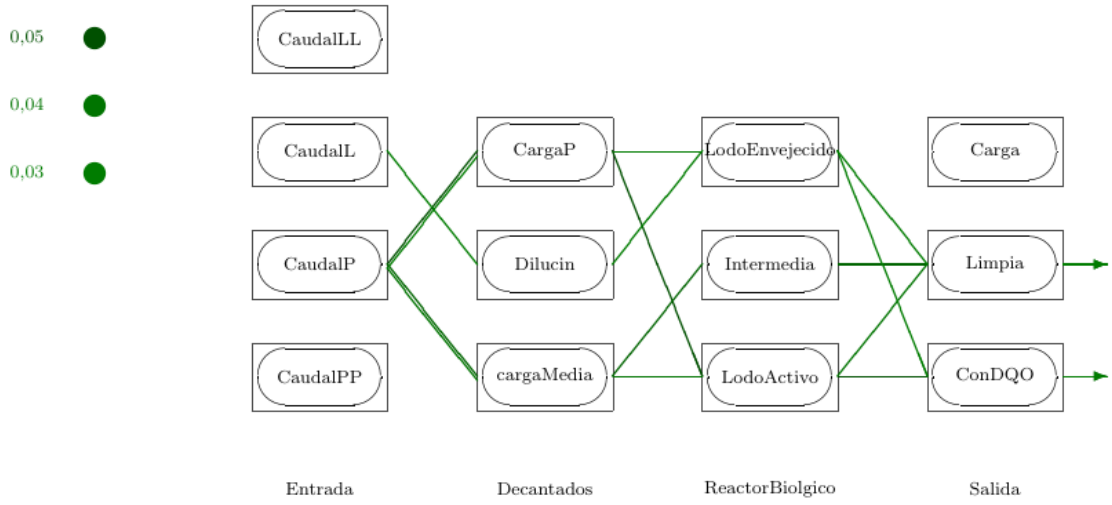


Figura 33 Diagrama de trayectoria con etiquetas

A partir de aquí es sencillo que cualquier experto comprenda que las trayectorias representadas están describiendo los siguientes estados de operación: Entrada, Decantados, ReactorBiológico, Salida

Con este diagrama de trayectorias podemos revisar fácilmente el comportamiento de los estados de operación de la planta representados a lo largo del proceso de depuración de agua. Esta herramienta es muy fácil de interpretar por parte de los expertos cuando se conoce el significado de cada clase, y los nodos del diagrama de operación están convenientemente etiquetados. Para esto es necesario la ayuda de un experto que colabore en la interpretación

A la vista del diagrama los expertos disponen de una herramienta interesante que identifica 5 regímenes de operación que determinan los siguientes escenarios:

T1: Representa una situación en la que llega bastante caudal a la planta y se trata de un agua bastante sucia que no se depura en decantación. El reactor biológico está trabajando a pleno rendimiento, pero aún y así no consigue eliminar toda la materia orgánica y los niveles de demanda química en el agua de salida de la planta no son todo lo bajos que pudieran ser

T2: Representa una situación en la que llega bastante caudal a la planta y se trata de un agua con niveles más moderados de contaminación que el caso anterior. El reactor biológico está trabajando a medio rendimiento y este nivel es suficiente para eliminar toda la materia orgánica arrojando al medio agua de alta calidad

T3: Representa una situación en la que llega bastante caudal a la planta y se trata de un agua con niveles más moderados de contaminación que el caso 1. El reactor biológico está trabajando a pleno rendimiento y elimina toda la materia orgánica arrojando al medio agua de alta calidad. En principio, y comparando con la anterior, esta parecería una situación más ineficiente, porque quizás reduciendo la aireación del tanque ya se podría realizar el tratamiento con óptimos resultados

T4: Representa una situación en la que llega bastante caudal a la planta y se trata de un agua bastante sucia que no se depura en decantación. El reactor biológico está trabajando a bajo rendimiento y así no consigue eliminar toda la materia orgánica y los niveles de demanda química en el agua de salida de la planta no son todo lo bajos que pudieran ser. Habría que estudiar por qué con los niveles de agua que entran no se aumentó la aireación del tanque, o si hay algún problema de funcionamiento

T5: Representa una situación extraña en que llega muy poco caudal a la planta, en situación de dilución. El lodo del reactor es antiguo y el agua sale bastante limpia. A juzgar por los expertos esto alude a una situación en que está bloqueado el biorreactor, y se ha bypasado el agua (muy diluida) directamente a salida. Tras algunas comprobaciones se ha podido verificar que esta trayectoria representa las situaciones de tormenta, donde el

caudal aumenta tanto que rompería el equilibrio biológico del biorreactor. Como la contaminación se diluye de forma natural por motivo de la propia lluvia, el agua se puede enviar directamente al medio sin tratar y lo que se hace en estas situaciones es cerrar las válvulas, lo que envejece el lodo biológico del biorreactor incrementando la edad media del mismo.

3.2 Caso de estudio OMS

Para la presentación de este caso utilizaremos los datos del estudio proporcionados por "WHO AIMS v2.2", que forman parte del estudio realizado para la OMS sobre la Identificación de perfiles de los sistemas de salud Mental de países LAMIC (Low And Middle – Income Countries) [Gibert 2009].

Este proyecto fue dirigido por la Dra. Karina Gibert directora de esta Tesis, y se hizo uso de Klass para la clasificación basada en reglas la adquisición de termómetros proporcionados por los expertos y la generación de cuadros semáforo. Para llegar al objetivo de la OMS de perfilar la situación de los países LAMIC ante la falta de información tan importante, y priorizar la atención de salud mental en los países LAMIC, así como definir los planes de intervención en los distintos tipos de países.

Este dataset está conformado por la información recogida de un conjunto de 42 países seleccionados como LAMIC, compuesto de 22 facetas, 155 ítems y 256 variables, y considerando los siguientes dominios:

- Política y marco legislativo
- Servicios de salud mental
- Salud mental en atención primaria
- Recursos humanos
- Información pública y vínculos con otros sectores
- Monitoreo e investigación
- Además de indicadores sintéticos dados por la OMS y el Banco Mundial

En ese proyecto se utilizó Clasificación Basada en reglas con conocimiento a priori de los expertos sobre el ámbito de estudio, si bien este detalle queda fuera del alcance de esta tesis.

3.2.1 Definición dataset

Para el presente caso el dataset cuenta con 256 variables, de las que tomaremos en cuenta 23variables relevantes que para el desarrollo de este caso.

A continuación, una descripción de las variables seleccionadas:

Variable	Descripción
Recursos del sistema -RS	
Incgroup	Nivel de ingreso del país.
Totprofmh	Número total de profesionales dedicados a la salud mental en el país por cada 100 000 habitantes.
Usmhexperca	Gasto en salud mental per cápita en USD.
Treatpre	Parte de la población diagnosticada y atendida por cada 100 000 habitantes.
Capratiosch	Cobertura del tratamiento de la esquizofrenia.
Outpfrate	Instalaciones ambulatorias por 100 000 habitantes.
Daytrfrate	Instalaciones de centros de día por 100 000 habitantes.
D4F1i11psychi	Número de psiquiatras
D4F1i12doctors	Número de doctores
D4F1i13nurses	Número de enfermeras
D4F1i14psycho	Número de Psicólogos

D4F1i15socwork	Número de trabajadores social
Estado Aproximación de la salud mental - SM	
d2f11i1closepsybeds	Camas psiquiátricas ubicadas en o cerca de la ciudad más grande (proporción per cápita).
d1f5i2exmhos	Gasto en hospitales mentales (%).
d2f6i71mhrec10y	Proporción de pacientes que permanecen en hospitales psiquiátricos durante 10 años o más.
Comcarewor	Proporción de usuarios tratados en hospitales mentales.
lundpararectrail	Ratio entre consultas externas y días en que el paciente está hospitalizado, indica si el sistema de salud da prioridad a mantener al paciente o ingresarlo lo más pronto posible.
Cbusrate	Unidades de pacientes internados basadas en la comunidad por 100 000 habitantes.
Estado Marco Legislativo - ML	
(D5F2i51)relprimcare	Relación de colaboración formal con el departamento de atención primaria.
D3f1i3Manuals	Disponibilidad de manuales de tratamiento y evaluación en atención primaria.
Legisl	Presencia de legislación.
Polplanr	Presencia de un plan de salud mental.
d6f1i6govmhrep	Informe sobre salud mental publicado por el departamento de salud del gobierno.

Tabla 14 Dataset OMS

Realizamos la carga de los datos del dataset y definimos las 23 variables activas que usaremos para el desarrollo de este caso.

3.2.2 Definición de procesos

En este caso se utiliza la Clasificación por estados como una implementación base para realizar multiview clustering, y esto es posible identificando las vistas de la base de datos con los estados del proceso. Para definir las vistas y el proceso que le asociamos será necesario agrupar las variables activas en los estados que serán asignados al proceso, se ha definido 3 estados:

Estado 1: Recursos del sistema RS, que contiene las variables: Incgroup, Totprofmh, Usmhexperca, Treatpre, Capratiosch, Outpfrate, Daytrfrate, D4F1i11psychi, D4F1i12doctors, D4F1i13nurses, D4F1i14psycho, D4F1i15socwork

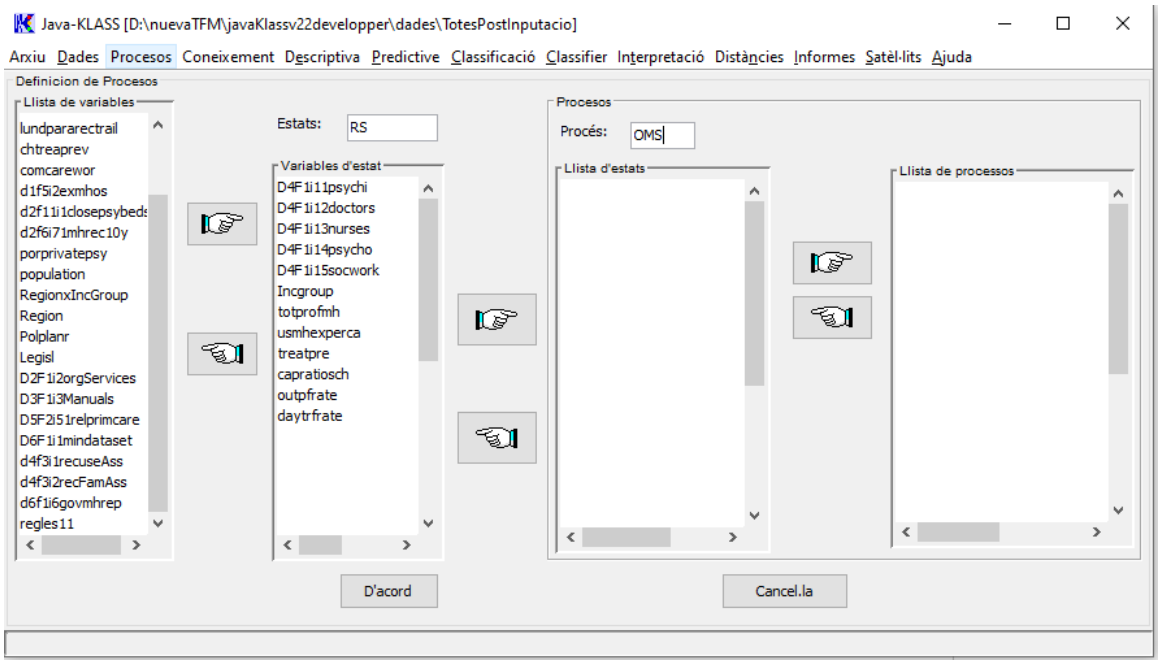


Figura 34 Definición Proceso RS

Estado 2: Aproximación de la atención en la salud mental (más concentrada en la inclusión social del paciente mental o en el tratamiento en hospitales) - SM, que contiene las variables: d2f11i1closepsybeds, d1f5i2exmhos, d2f6i71mhrec10y, Comcarewor, lundpararectrail, Cbusrate.

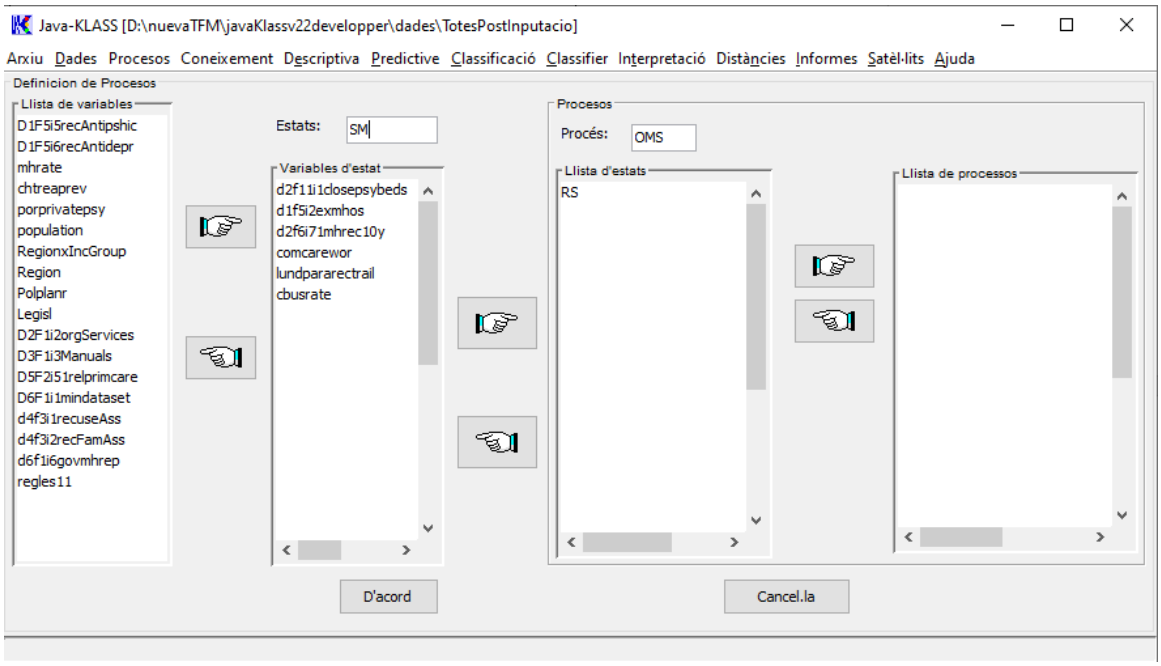


Figura 35 Definición estado SM

Estado 3: Marco Legislativo - ML, que contiene las variables: (D5F2i51)relprimcare, D3f1i3Manuals, Legisl, Polplanr, d6f1i6govmhrep.

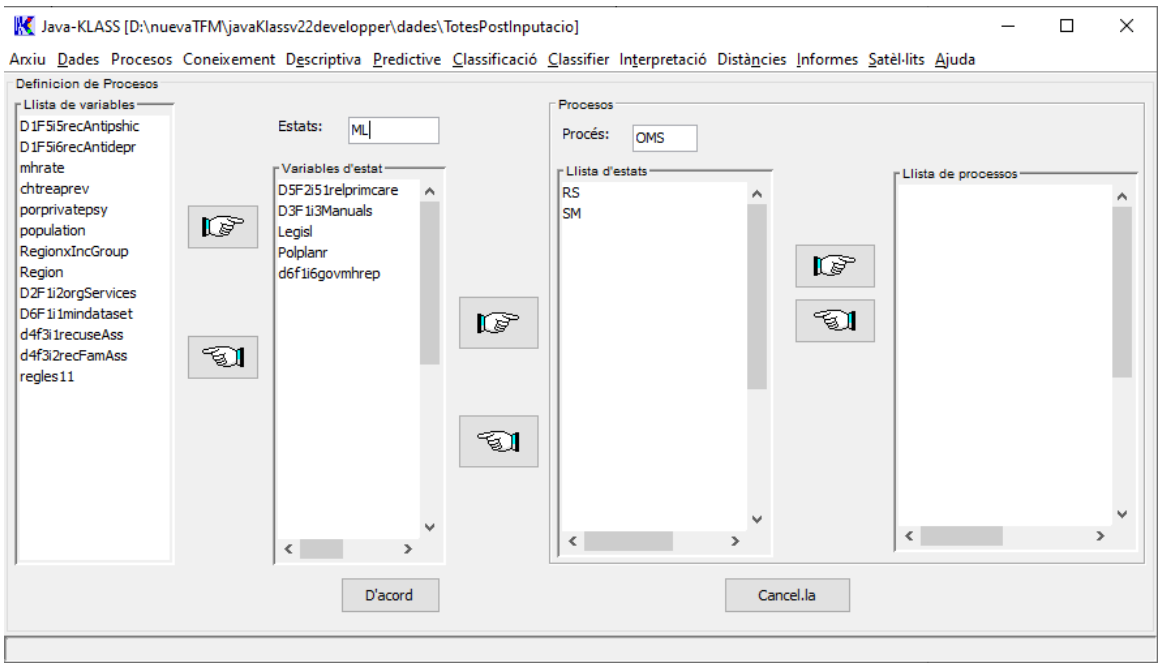


Figura 36 Definición estado ML

Creación de proceso denominado OMS que contendrá los estados creados anteriormente.

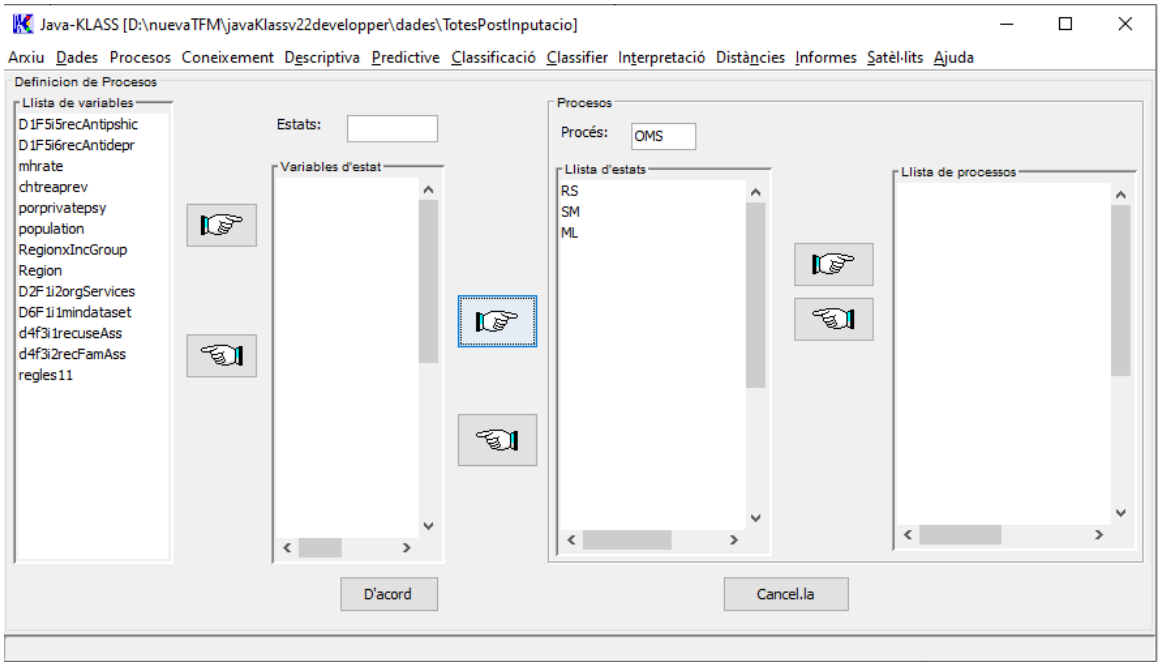


Figura 37 Definición proceso OMS

3.2.3 Clasificaciones basadas en reglas por estados

Para la clasificación basada en reglas por estados que se les aplicará a los estados del proceso creado se usaran los siguientes parámetros: Criterio de agregación de Ward, distancia Mixta de Gibert automático, calculara $\alpha = 0.032588556$ y $\beta = 0.967441146$

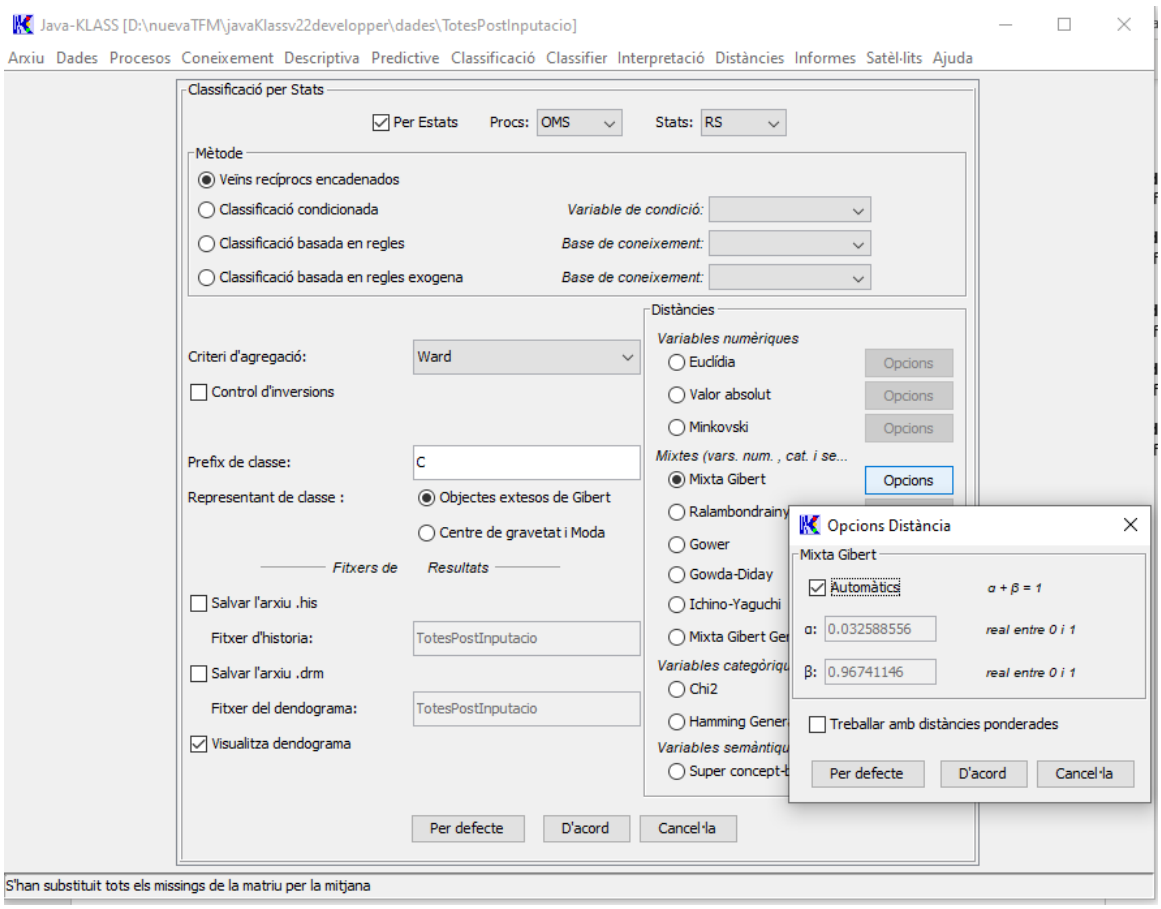


Figura 38 Clasificación por estados proceso OMS

Realizada la clasificación se generarán los dendrogramas por cada estado procesado, a los cuales hay que analizar y realizar el podado, detallamos a continuación.

Estado 1: Recursos del sistema RS, se realizará el corte en 3 clases del dendrograma generado de la clasificación

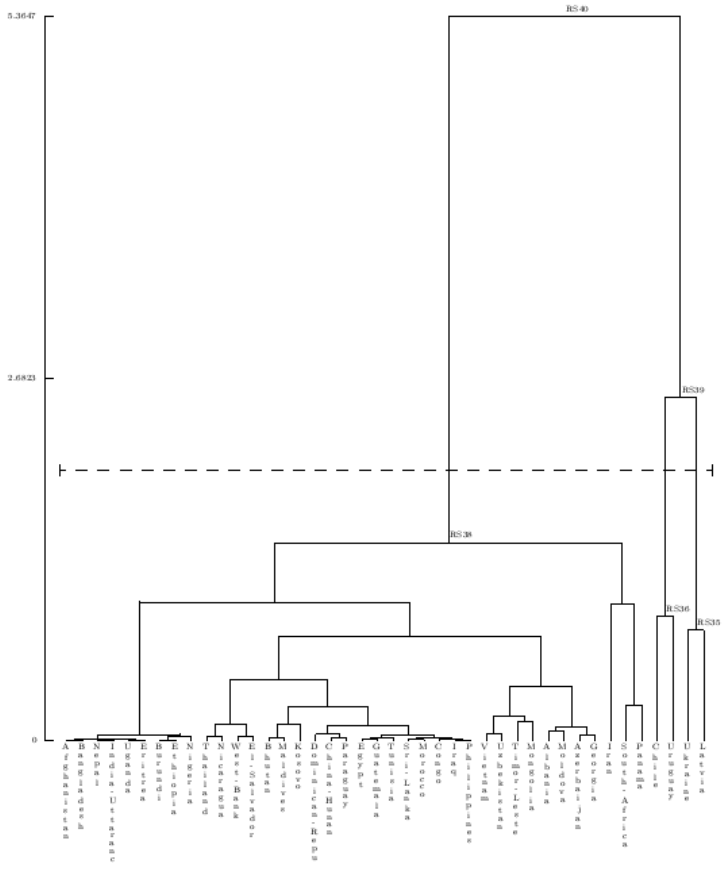


Figura 39 Corte dendrograma RS

Generamos la descriptiva para visualizar el comportamiento de las variables que conforman el estado., para mayor información se puede ver el Anexo 6

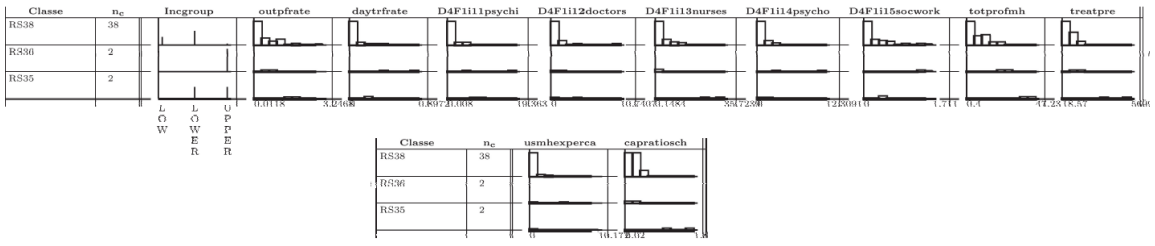


Figura 40 Panel de clases estado RS

	Etiqueta	Incgroup	Outpfrate	Daytrfrate	D4F1i11psychi	D4F1i12doctors	D4F1i13nurses	D4F1i14psycho	D4F1i15socwork	Totprofmh	Treatpre	Usmhexperca	Capratiosch
RS38	Scarcity	Lower/low	BAJO	BAJO	BAJO	BAJO	BAJO	BAJO	BAJO	BAJO	BAJO	BAJO	BAJO
RS36	"+Staff/+USMH"	Upper	MEDIO	ALTO	ALTO	BAJO	BAJO	ALTO	ALTO	MEDIO	BAJO	MEDIO	BAJO
RS35	"+Nurses+TreatPre"	Low/Uppe	ALTO	MEDIO	MEDIO	MEDIO	ALTO	BAJO	MEDIO	ALTO	ALTO	BAJO	ALTO

Tabla 15 Conceptualización de las clases RS del proceso OMS

Estado 2: Estado Aproximación de la salud mental – SM, se realizará el corte en 4 clases del dendrograma generado de la clasificación

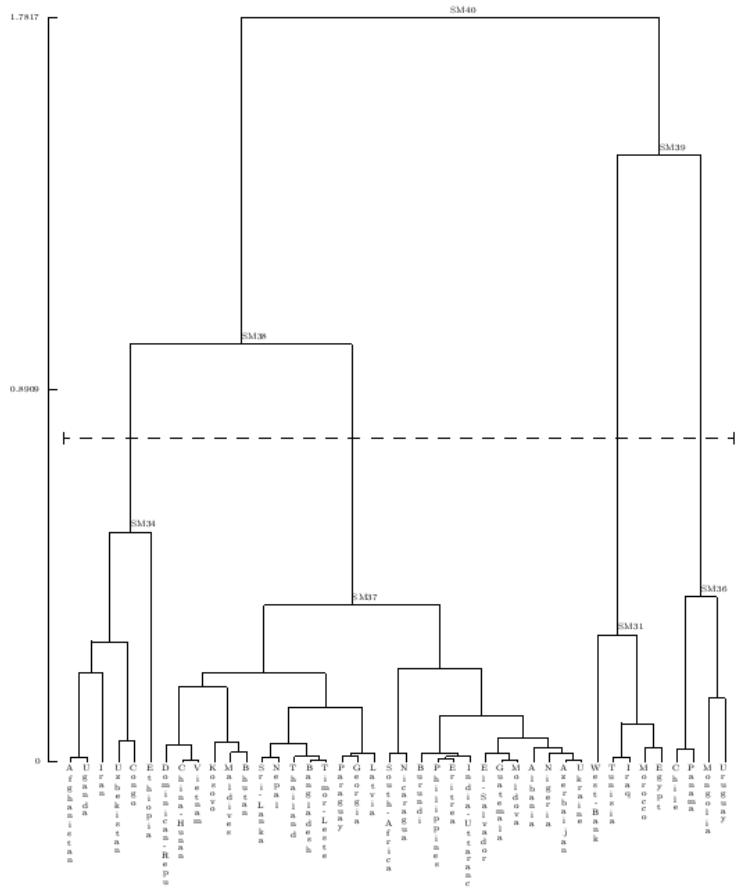


Figura 41 Corte Dendrograma SM

Generamos la descriptiva para visualizar el comportamiento de las variables que conforman el estado, para mayor información se puede ver el Anexo 7

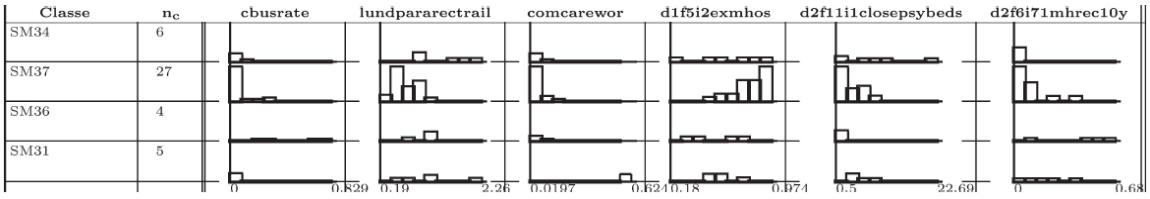


Figura 42 Panel de clase estado SM

	Etiqueta	Cbusrate	lundparare ctrail	Comcarew or	d1f5i2exm hos	d2f11i1clos epsybeds	d2f6i71mhr ec10y
SM34	communityCent	BAJO	ALTO	BAJO	MEDIO+	ALTO	BAJO
SM37	hospCent+Res	BAJO	BAJO	BAJO	ALTO	MEDIO	MEDIO
SM36	hospCent	MEDIO	MEDIO	BAJO	MEDIO+	BAJO	ALTO
SM31	Transition	BAJO	MEDIO+	ALTO	MEDIO	MEDIO+	MEDIO+

Tabla 16 Conceptualización de las clases SM del proceso OMS

Estado 3: Estado Marco Legislativo – ML se realizará el corte en 3 clases del dendrograma generado de la clasificación

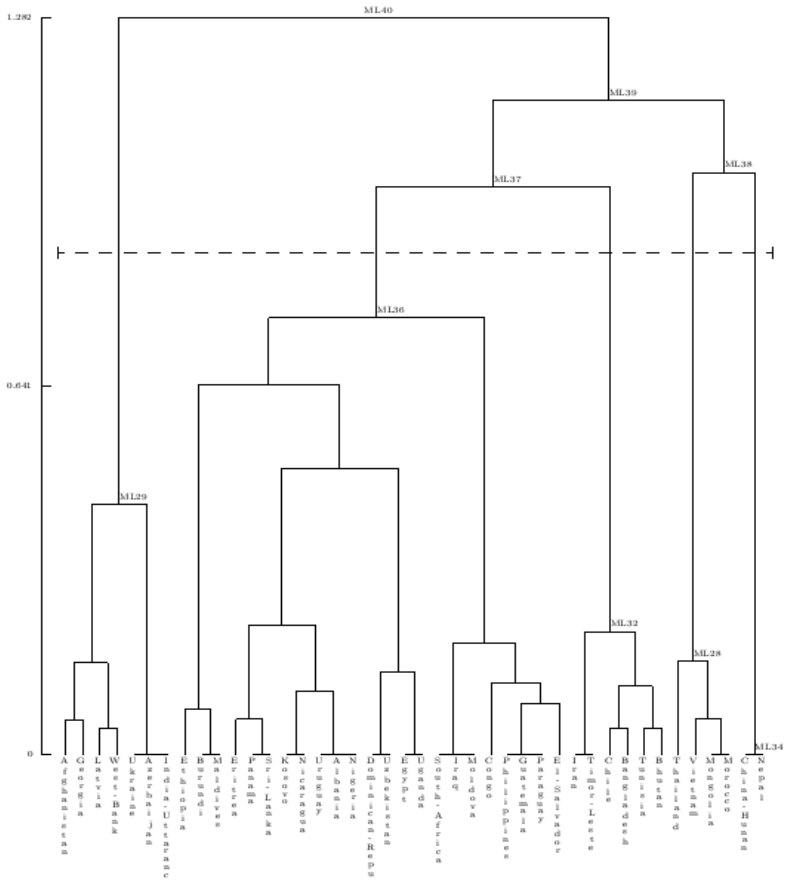


Figura 43 Corte Dendrograma del estado ML

Generamos la descriptiva para visualizar el comportamiento de las variables que conforman el estado, para más información se puede ver el Anexo 8

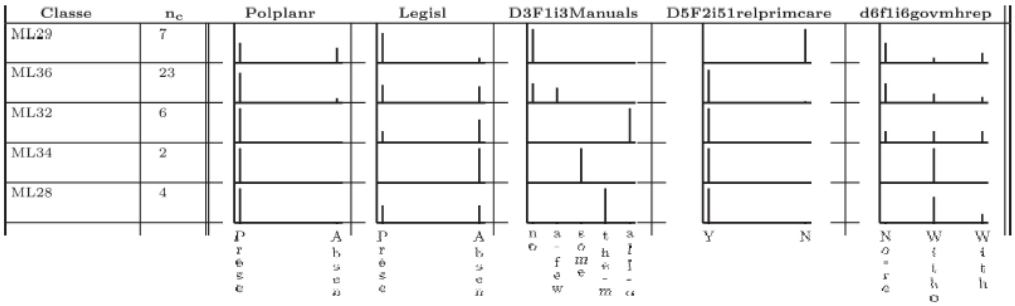


Figura 44 Panel de clases estado ML

	Etiquetas	Polplanr	Legisl	D3f1i3Manuals	D5F2i51relprimcare	d6f1i6govmhrep
ML29	Scarce	Half	Most	No	No	No
ML36	in transition	Most	Half	Few	Y	Few
ML32	Developped-	All	some	All	Y	heterogeneity
ML34	No legal framework	All	No	Some	Y	rep WithoutMH
ML28	Developped+	All	Half	Most	Y	rep WithoutMH

Tabla 17 Conceptualización de las clases ML del proceso OMS

3.2.4 Generación de diagrama de trayectoria

Luego de realizar el corte de los dendrogramas, podemos realizar la generación del diagrama de trayectorias con las variables de clases que se crearon, se realizará el cálculo de la frecuencia de la nueva variable de trayectoria que contendrá la concatenación de las variables de clases.

Revisando la Tabla de trayectoria a continuación, se observa que se ha generado 17 trayectorias posibles.

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
RS38+SM34+ML29	1	1	0.0238	0.0238
RS38+SM37+ML36	15	16	0.3571	0.381
RS38+SM37+ML29	3	19	0.0714	0.4524
RS38+SM37+ML32	3	22	0.0714	0.5238
RS36+SM36+ML32	1	23	0.0238	0.5476
RS38+SM37+ML34	2	25	0.0476	0.5952
RS38+SM34+ML36	4	29	0.0952	0.6905
RS38+SM31+ML36	2	31	0.0476	0.7381
RS38+SM34+ML32	1	32	0.0238	0.7619
RS35+SM37+ML29	2	34	0.0476	0.8095
RS38+SM36+ML28	1	35	0.0238	0.8333
RS38+SM31+ML28	1	36	0.0238	0.8571
RS38+SM36+ML36	1	37	0.0238	0.881
RS38+SM37+ML28	2	39	0.0476	0.9286
RS38+SM31+ML32	1	40	0.0238	0.9524
dades mancants	0	N = 42	0	

Tabla 18 Tabla de frecuencia variable de trayectoria Proceso OMS

Haciendo uso de la opción panel trayectoria implementada en esta tesis de Máster, y seleccionando las variables de clase de cada estado, generaremos el diagrama de trayectoria con las 8 trayectorias más frecuentes según la tabla de frecuencia.

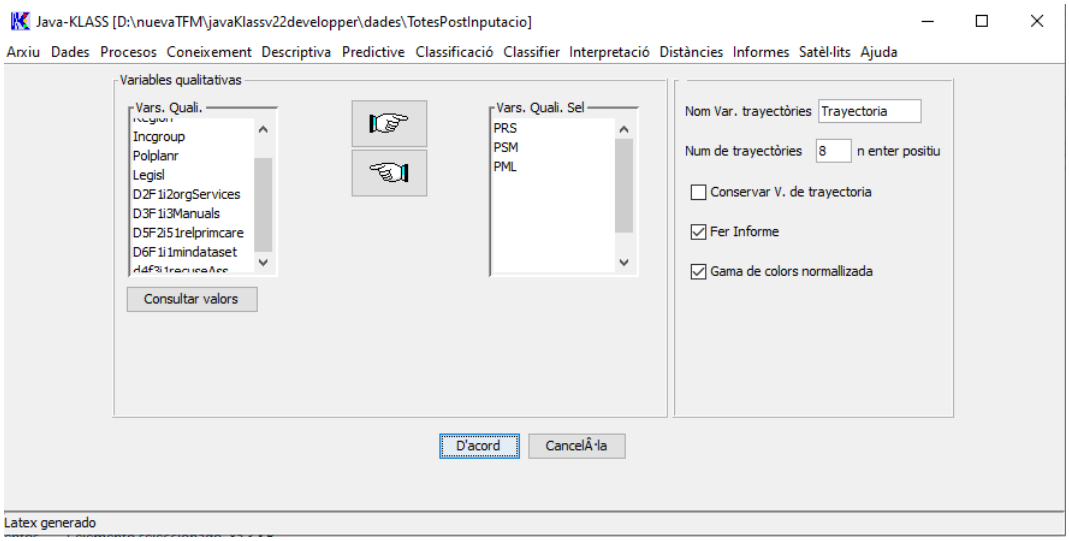


Figura 45 Panel Diagrama trayectoria

Obtenemos como resultado el diagrama mostrado a continuación

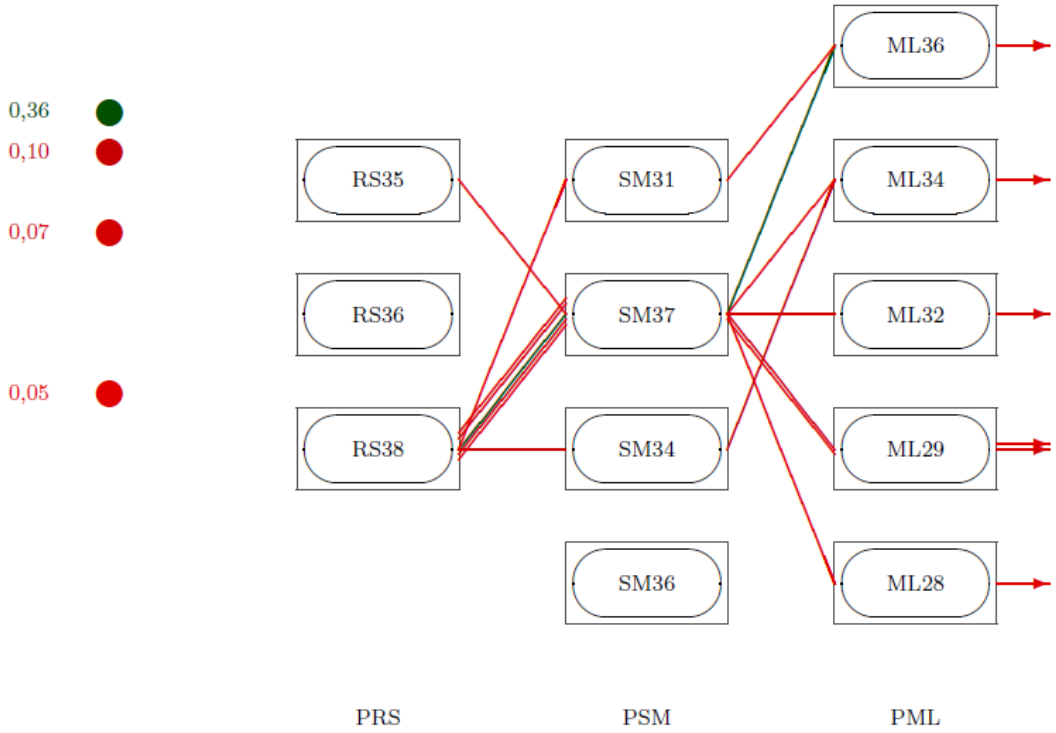


Figura 46 Diagrama Trayectoria Proceso OMS (8 trayectorias más frecuentes)

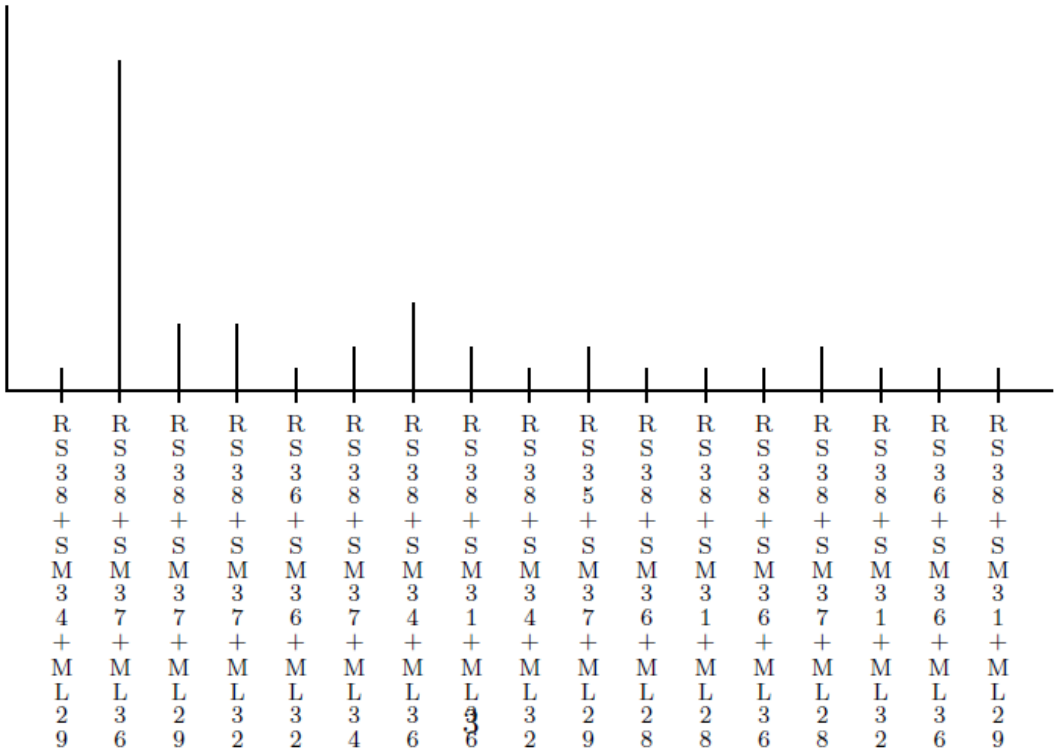


Figura 47 Diagrama de Barra OMS

Podemos Observar que la trayectoria RS38+SM37+ML36, es la más frecuente con 0.36 de probabilidad
La segunda más frecuente es la trayectoria RS38+SM34+ML36 con el 0.10 de probabilidad
Las trayectorias que tienen el 0.07 de probabilidad son 2:
RS38+SM37+ML29
RS38+SM37+ML32
Las trayectorias que tienen el 0.05 de probabilidad son 4:
RS38+SM37+ML28
RS38+SM37+ML34
RS35+SM37+ML29
RS38+SM31+ML36

Podemos observar que el nodo SM37 es el nodo que más comparten las trayectorias representadas
Esta información gráfica servirá de ayuda para los expertos para la interpretación de los resultados.
Para mayor información se puede ver el Anexo 9, con el reporte del diagrama de trayectoria del proceso OMS

Si este diagrama se realiza con las clases etiquetadas por los expertos su potencia expresiva es mucho mayor.

Taula de freqüències				
Modalitats	**Freq. absol.**	**Freq. acum.**	**Freq. relat.**	**Freq. rel. acum.**
Scarcity+communityCent+Scarce	1	1	0.0238	0.0238
Scarcity+HospCebtPRes+In transation	15	16	0.3571	0.381
Scarcity+HospCebtPRes+Scarce	3	19	0.0714	0.4524
Scarcity+HospCebtPRes+DevelopedL	3	22	0.0714	0.5238
PStaff/PUSMH+HospCent+DevelopedL	1	23	0.0238	0.5476
Scarcity+HospCebtPRes+No Legal Framework	2	25	0.0476	0.5952
Scarcity+communityCent+In transation	4	29	0.0952	0.6905
Scarcity+Transition+In transation	2	31	0.0476	0.7381
Scarcity+communityCent+DevelopedL	1	32	0.0238	0.7619
PNUrsesPTreatPre+HospCebtPRes+Scarce	2	34	0.0476	0.8095
Scarcity+HospCent+DevelopedP	1	35	0.0238	0.8333
Scarcity+Transition+DevelopedP	1	36	0.0238	0.8571
Scarcity+HospCent+In transation	1	37	0.0238	0.881
Scarcity+HospCebtPRes+DevelopedP	2	39	0.0476	0.9286
Scarcity+Transition+DevelopedL	1	40	0.0238	0.9524
PStaff/PUSMH+HospCent+In transation	1	41	0.0238	0.9762
Scarcity+Transition+Scarce	1	42	0.0238	1
dades mancants	0	N = 42	0	

Tabla 19 Tabla de frecuencia variable de trayectoria con etiquetas Proceso OMS

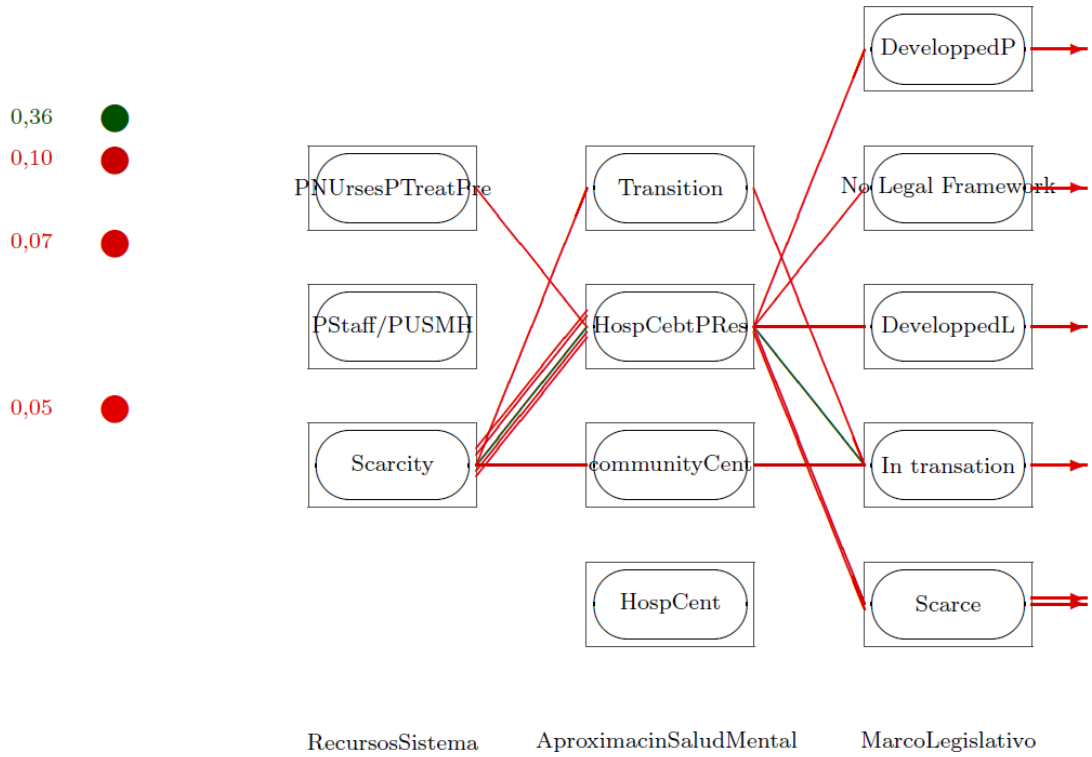


Figura 48 Diagrama Trayectoria Proceso OMS (8 trayectorias más frecuentes) con Etiquetas

Aquí se observan 8 trayectorias que representan los siguientes tipos de países:

T1 países con gran falta de recursos orientación de la salud mental bastante orientada al ingreso en reclusión hospitalaria y con un marco legal en transición. Ya existe conexión con atención primaria pero no hay marco legal

T2: países con gran falta de recursos orientación de la salud mental centrada en la nueva aproximación de inclusión social y tratamiento orientado a la comunidad. Marco legal en transición también

T3: países con gran falta de recursos orientación de la salud mental bastante orientada al ingreso en reclusión hospitalaria y marco legal muy precario

T4: países con gran falta de recursos orientación de la salud mental bastante orientada al ingreso en reclusión hospitalaria y cierto desarrollo del marco legal

T5: países con gran falta de recursos orientación de la salud mental bastante orientada al ingreso en reclusión hospitalaria y con los marcos legales más desarrollados del estudio

T6: países con gran falta de recursos orientación de la salud mental bastante orientada al ingreso en reclusión hospitalaria y ausencia de marco legal

T7: Países con mayor concentración de enfermeras y mayor capacidad de cobertura de la población con problemas mentales, Orientación del servicio centrada en el hospital y marco legal precario

T8: Países con gran falta de recursos, orientación del servicio en transición entre la atención centrada en el hospital y la atención en trabajos con la comunidad y marco legal también en transición.

Capítulo 4

Conclusiones

4.1 Conclusiones

Esta tesis de máster está enfocada en la representación del conocimiento dinámico del proceso de una forma gráfica añadiendo una nueva funcionalidad al sistema KLASS. El diagrama de trayectorias, una representación gráfica que nos indica los patrones más frecuentes observados a lo largo de un proceso. Se ha ilustrado cómo la herramienta es también útil para analizar constructos complejos compuestos de varias dimensiones

Como punto de partida ha sido necesario adquirir conocimientos en las herramientas y metodologías que se han utilizado en la implementación de esta tesis, conocer conceptos de datamining. Ha sido muy importante para las diferentes etapas de pruebas, de interpretación de datos y de representación de resultados, así como de las herramientas necesarias para su implementación, tales como LaTEX que sirve para la generación de reporte, temas de clustering, de postprocesado de sus resultados, etc

Se presenta una metodologia para el análisis de trayectorias en datos de procesos discretos, así como el uso de algoritmos de identificación de trayectorias para presentar las trayectorias más frecuentes, así como fue importante como primer paso entender el concepto de proceso y estados, para así optimizar la forma actual que se manejaba la definición de los procesos, y de esa forma realizar la Clasificación Basada en Reglas por Estados de una forma correcta.

En el caso de estudio (Aguas residuales) partiendo de un conocimiento a priori los investigadores definieron un proceso con 4 estados: Entrada, Decantador, Bioreactor y Salida. Posteriormente se realizó la clasificación basada en reglas por estados, el cual generó automáticamente 4 dendrogramas correspondiente al proceso. Y se ha podido ver cómo la interpretación de trayectorias arroja luz sobre estados ineficientes de operación de la planta o situaciones excepcionales como los días de tormenta. En el caso de estudio de

la OMS se utiliza la herramienta para analizar el constructo Sistemas de Salud Mental en función de tres dimensiones Recursos, Orientación del Servicio, Marco Legal y se observan patrones de países en distintos estados de evolución entre la atención mental centrada en el hospital o la comunidad (aproximación moderna), entre la ausencia de marco legal o un mayor desarrollo y entre distintos niveles de recursos, que arrojan información relevante a los expertos sobre cómo potenciar el desarrollo de la Salud Mental en dichos países.

Si bien KLASS ofrece los paneles de clase y las descriptivas por clase para identificar los patrones de una clasificación, sin embargo, no existía todavía en KLASS una forma visual para ilustrar cómo se relacionan los comportamientos de los distintos estados entre sí. Por lo tanto, el desarrollo de esta tesis de master está enfocada en realizar la representación gráfica de los estados del proceso y visualizar la secuencia, utilizando el concepto de diagrama de trayectorias ya desarrollado en el propio grupo de investigación.

En realidad, el concepto de diagrama de trayectorias se puede utilizar para representar interacciones entre cualquier conjunto de variables cualitativas. Y el mecanismo se ha implementado en el caso general y desligado de la clasificación por estados, de forma que se pueda utilizar desde cualquier otro contexto en una sesión de KLASS. Las modalidades de cada una de las etapas del proceso son representadas por nodos, y los arcos que la conectan como **(arco de frecuencia).** Los niveles de importancia de las trayectorias se han asociado a sus frecuencias de ocurrencia (que según una aproximación frecuentista de la probabilidad representan estimaciones de cuán probables son) y se ha asociado dicha frecuencia a un modelo de color donde las trayectorias más frecuentes se pintan en verde oscuro y van bajando al verde claro, amarillo oscuro, amarillo brillante, rojo oscuro y rojo brillante para las más raras. De esta forma se describe el comportamiento del flujo del proceso y su estado de forma visual, se pueden identificar las variables que determinan cierto comportamiento dentro del proceso ejecutado.
Con la incorporación de esta nueva funcionalidad a KLASS, se la dota de una nueva forma de interpretación que complementa las ya implementadas, siendo éste un método para representar patrones dinámicos que permitirá la comprensión de los fenómenos por parte del experto para poder apoyar la toma de decisiones complejas. Este tipo de representación gráfica es muy útil para las tomas de decisiones en distintos ámbitos desde empresas a gobiernos es por eso de su importancia

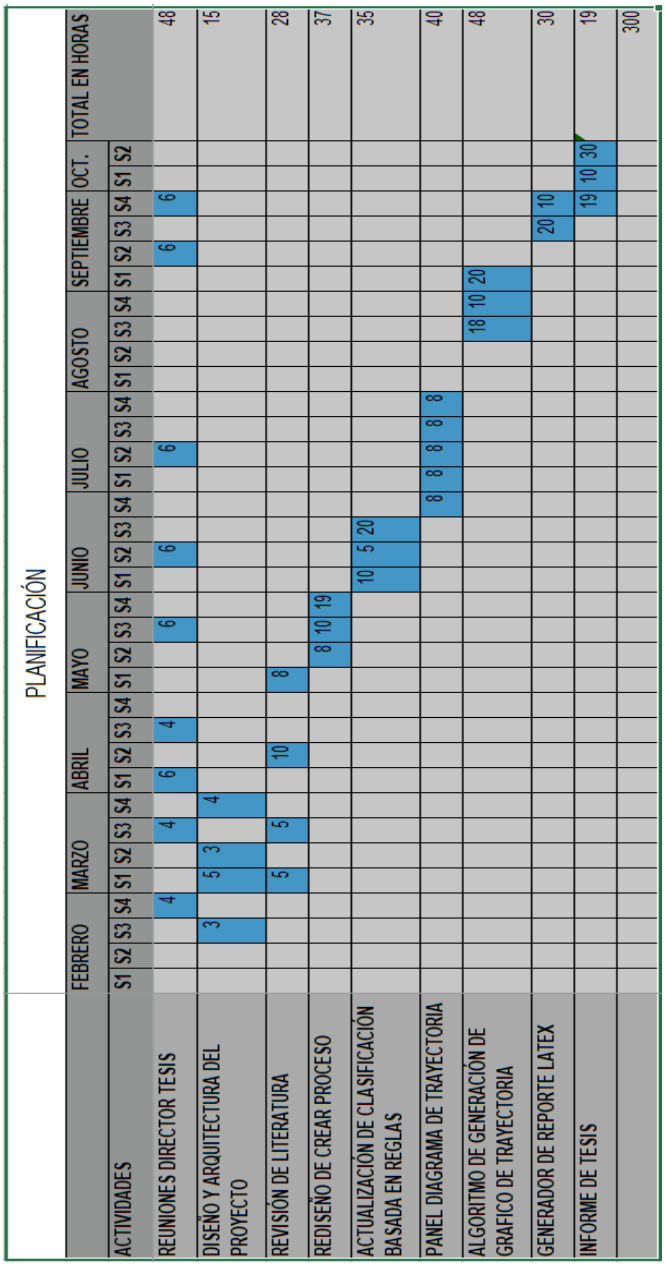
4.2 Recomendaciones y trabajo futuro

La clasificación basada en reglas de estado genera N dendrogramas de acuerdo con número de estados que tenga el proceso, para lo cual el usuario deberá acceder a cada uno de ellos y decidir los respectivos niveles de corte que generan las variables de estado que se utilizan en el diagrama de trayectorias. En este momento este paso se realiza manualmente a partir de la componente ya existente en KLASS que permite cortar dendrogramas y generar las variables de clase a partir del número de clases que le indica el usuario. Para trabajos futuros, debería conectarse esta componente con los dendrogramas que produce el propio proceso de Clasificación basada en reglas por estados y generar rápidamente las variables de clase. Ello permitirá la conexión automática con el nuevo módulo de construcción del diagrama de trayectorias y la conexión del informe resultado al informe del propio proceso de clasificación. Sin embargo, ello requiere la implementación de criterios automáticos de corte del dendrograma que son actualmente objeto de desarrollo en el grupo, puesto que KLASS realiza clasificaciones con variables de distintos tipos simultáneamente y la literatura aporta criterios solamente para matrices de variables numéricas.

Capítulo 5

Planificación y costos

5.1 Planificación



5.2 Costos

	PRESUPUESTO									
GASTOS	FEBRERO	MARZO	ABRIL	MAYO	JUNIO	JULIO	AGOSTO	SEPTIEMBRE	OCTUBRE	TOTAL
TRANSPORTE	€ 50,00	€ 50,00	€ 50,00	€ 50,00	€ 50,00	€ 50,00	€ 50,00	€ 50,00	€ 50,00	€ 450,00
EQUIPO DE COMPUTACIÓN	€ 1200,00	€ 0,00	€ 0,00	€ 0,00	€ 1200,00	€ 0,00	€ 0,00	€ 0,00	€ 0,00	€ 2400,00
PROGRAMADOR	€ 800,00	€ 800,00	€ 800,00	€ 800,00	€ 800,00	€ 800,00	€ 800,00	€ 800,00	€ 800,00	€ 7200,00
ADMINISTRADOR DE PROYECTO	€ 1800,00	€ 1800,00	€ 1800,00	€ 1800,00	€ 1800,00	€ 1800,00	€ 1800,00	€ 1800,00	€ 1800,00	€ 16200,00
TOTAL POR MES	€ 3850,00	€ 2650,00	€ 2650,00	€ 2650,00	€ 3850,00	€ 2650,00	€ 2650,00	€ 2650,00	€ 2650,00	€ 26250,00
COSTO TOTAL DEL PROYECTO	€ 52500,00									

Glosario

Glosario

CIBR	*Clasificación basada en reglas*
CIBRxE	*Clasificación basada en reglas por estados*
BC	*Base de conocimientos*
TLP	*Traffic Ligth Panel*
LAMIC	*Low And Middle – Income Countries*
OMS	*Organización Mundial de la Salud*

Bibliografía

Bibliografía

[Benzecri 1973] Benzecri JP, L'analyse des donnés. Tome 1: La Toxinomie, Tome 2: L'analyse des correspondences. 1ª Ed 1973. Paris Dunod, 1980.

[Calinski and Harabasz, 1974] Calinski, T. and Harabasz, J. (1974) A dendrite method for cluster analysis. Communications in Statistics - Simulation and Computation, 3(1):1—27, Taylor and Francis

[De los Reyes, 2019] De los Reyes, J. (2019). Implementación de un módulo avanzado de imputación de datos faltantes para KLASS. Master's thesis, Master's thesis, UPC.

[García Rudolph, 2009] García Rudolph, A. (2009). Desarrollo de una metodología de extracción de conocimiento (KDD) en patrones dinámicos: evolución de la calidad de vida en pacientes con afectación neurológica. PhD thesis, UPC.

[Gibert, 1991] Gibert, K. (1991). Klass. estudi d'un sistema d'ajuda al tractament estadístic de grans bases de dades. Master's thesis, Master's thesis, UPC.

[Gibert, 1995] Gibert, K. (1995). L'us de la informació simbólica en l'automatització del tractament estadístic de dominis poc estructurats. PhD thesis.

[Gibert, 2014] Gibert K 2014 Automatic generation of classes interpretation as a bridge between clustering and decision making International Journal of Multicriteria Decision Making 4(2):154-182 Inderscience

[Gibert et al., 2012b] Gibert, K., Conti, D., and Vrecko, D. (2012b). Assisting the end-user in the interpretation of proles for decision support. an application to wastewater treatment plants. Environmental Engineering and Management Journal, 11(5):931{944.

[Gibert et al., 2013] Gibert, K., Rodríguez-Silva, G., and Annicchiarico, R. (2013). Post-processing: Bridging the gap between modelling and effective decision-support. the profile

assessment grid in human behaviour. Mathematical and Computer Modelling, 57(7-8):1633-1639.

[Gibert et al., 2008a] Gibert, K., García-Rudolph, A., and Rodrìguez-Silva, G. (2008a). The role of kdd support-interpretation tools in the conceptualization of medical profiles: An application to neurorehabilitation. Acta Informatica Medica,16(4):178.

[Gibert et alt 2005] Gibert K, Nonell R; Velarde JM, Colillas MM. (2005). Knowledge discovery with clustering: impact of metrics and reporting phase by using KLASS. Neural Network World, 4: 319-326. ISSN: 1210-0552

[Gibert et al 2018] Gibert, K., J. Horsburgh, I. Athanasiadis, G. Holmes (2018) "Environmental Data Science." Environmental Modelling & Software 106: 4-12. (DOI:10.1016/j.envsoft.2018.04.005)

[Gibert, Conti 2015] Gibert K, D. Conti (2015) aTLP: A color-based model of uncertainty to evaluate the risk of decisions based on prototypes. Artificial Intelligence Communications 28:113-126, IOSPress

[Gibert, 2009]Gibert, K., Martin, J. C., and Salvador-Carulla, L. (2009). Who-aims: First analysis for missing values inputation. World Health Organization. Geneva. Switzerland.

[Gibert et al., 2010a] Gibert, K., García-Alonso, C., and Salvador-Carulla, L. (2010a). Integrating clinicians, knowledge and data: expert-based cooperative analysis in healthcare decision support. Health research policy and systems.

[Gibert Cortes 1997] Gibert, K., and Cortés, U. (1997). "Weighing quantitative and qualitative variables in clustering methods." Mathware and Soft Computing, 4(3), 251-266.

[Gibert and Conti, 2016] Gibert, K., Conti, D. (2016). On the understanding of profiles by means of post-processing techniques: an application to financial assets. Environmental Engineering and Management. International Journal of Computer Mathematics

[Gibert & Cortés 1998] Gibert K. Cortés U. Clustering based on rules and Knowledge Discovery in ill-structured domains. Computación y Sistemas 1(4) pp 213-227. ISSN 1405-5546 1998

[Gibert, García-Rudolph, Curcoll, Soler, Pla, Tormos 2009] K. Gibert, A. García-Rudolph, L. Curcoll, D. Soler, L. Pla, J. M. Tormos (2009) Knowledge Discovery about Quality of Life changes of Spinal Cord Injury patients: Clustering based on rules by states. In Studies in Health Technology and Informatics, v150: 579—583,Aug 2009. IOSSPress

[Gibert, Rodríguez-Silva, Rodríguez-Roda, 2010] Gibert K, Rodríguez-Silva G, Rodríguez-Roda I, 2010: Knowledge Discovery with Clustering based on rules by States: A water treatment application. Environmental Modelling&Software 25:712-723 https://doi.org/10.1016/j.envsoft.2009.11.004

[Gibert, Sevilla-Villanueva, Sànchez-Marrè 2016] Gibert, K, B. Sevilla-Villanueva, M. Sànchez-Marrè (2016) The role of significance tests in consistent interpretation of nested partitions. Journal of Computational and Applied Mathematics, 292: 623-633, Elsevier, Amsterdam, NL (htpps://doi.org/10.1016/j.cam.2015.01.031)

[Jordan, 2019] Jordan, C. (2019). Interpretación automática de clases y resoluci ón de conflictos en bases de conocimiento para KLASS. Master's thesis, Master's thesis, UPC.

[Mandadapu, 2019] Mandadapu, L. (2019). Implementing new interpretation oriented tools in Klass to support decision making based on logistic Regression, bachelor thesis

[Mollá Santiago, 2014] Mollá Santiago, S. (2014). Generalització de mètodes de density-based clustering a dades mixtes.

[Rodríguez 2009], Rodríguez, G. 2009. Metodología de Inducción e Interpretación de Perfiles Dinámicos (I2DPro). PhD thesis, UPC.

[Tukey77] Tukey JW : Exploratory Data Analysis, Addison-Wesley, 1977

LaTeX/Picture. (2019, March 1). *Wikibooks, The Free Textbook Project*. Retrieved 20:23, October 14, 2019 from https://en.wikibooks.org/w/index.php?title=LaTeX/Picture&oldid=3520704.

Guia de LaTeX, https://sites.google.com/site/guiadelatex/home, consultada 15/09/2019

Anexos

Anexos

- Anexos 1: Reporte descriptiva perclases PE
- Anexos 2: Reporte descriptiva perclases PD
- Anexos 3: Reporte descriptiva perclases PB
- Anexos 4: Reporte descriptiva perclases PS
- Anexos 5: Reporte diagrama de trayectoria proceso planta
- Anexos 6: Reporte descriptiva perclases PRS
- Anexos 7: Reporte descriptiva perclases PSM
- Anexos 8: Reporte descriptiva perclases PML
- Anexos 9: Reporte diagrama de trayectoria proceso OMS